Importance Sampling as an Efficient Strategy for the Conformational Analysis of Flexible Molecules†

Noham Weinberg,*,‡,§ Manjinder Dhaliwal,§ Christopher Reilly,§ Essex Edwards,§ and Saul Wolfe*,‡

Department of Chemistry, Simon Fraser University, Burnaby, BC, Canada, and Department of Chemistry, University of the Fraser Valley, Abbotsford, BC, Canada

Received: June 19, 2008; Revised Manuscript Received: October 5, 2008

We propose an importance sampling technique for the conformational analysis of flexible molecules that bridges the gap between currently used uniform random search and random walk methods, and compare performances of these methods in their application to selected open-chain and closed-ring hydrocarbons. It is suggested that, if no information on conformational properties of a molecule is available, the optimum strategy of its conformational analysis should include a uniform random search followed by an importance sampling biased as prescribed by the uniform search and concluded by a random-walk or genetic algorithm routine to rectify the properties of the most important parts of the conformational space.

1. Introduction

Random search techniques[1] have become a tool of choice in the conformational analysis of flexible molecules because of the enormity of the conformational space, whose size scales exponentially with the number of flexible links in a molecular system. With few exceptions, the most important of which are genetic algorithms[2] and importance sampling,[3] these methods can be classified as either random walk or uniform random search techniques.[4] Each has advantages and disadvantages: random walk methods are fast, focused, and information-efficient, but are localized around the starting point and show a significant dependence on this choice. In addition, they are data-dependent and not suitable for parallel computing. Uniform random search methods are ideally suited for parallel computing and provide global coverage of the conformational space, but lack focus and are slow in identifying specific targets, such as low-energy minima. Genetic algorithms seem to combine the advantages of both techniques but do not preserve closed rings and so cannot be used for cyclic systems unless they operate on a set of reasonably close structures. Importance sampling is widely used in various other applications,[5] and if the proper bias for sampling is provided, should afford an efficient search technique suitable for the conformational analysis of open-chain and cyclic molecules. In this work we formulate such a technique, demonstrate its efficiency, and compare it to other conformational search methods.

2. Computational Details

Randomly generated conformations were fully optimized with the MM3 program[6] and at the AM1 level[7] with Gaussian03.[8] The nature of optimized structures was verified by frequency calculations. Geometries of the minimized conformations were compared and duplicates removed. All energies are reported with reference to the respective global minima.

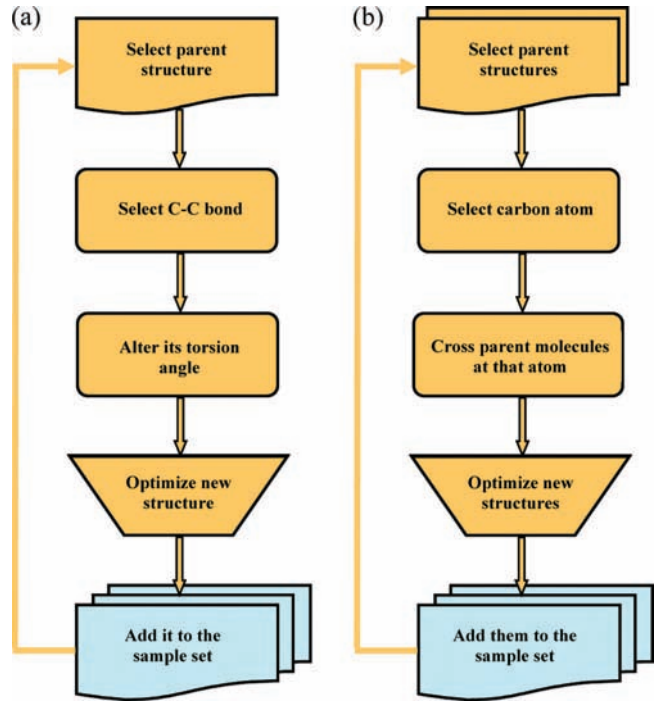

Figure 1. The flow charts of the simplified genetic algorithms with operations restricted to mutations only (a) or crossovers only (b).

For the comparison of different methods, sets of randomly generated structures were prepared by using the importance sampling method described below, our earlier reported version of the uniform random search,[4] the corner flap method,[9] and two simplified genetic algorithms (Figure 1) with operations restricted to mutations only or crossovers only.

To ensure that global minima have been found, comparisons were limited to 12-, 15-, and 18-membered hydrocarbon chains (with *all-anti* as global minima[10]) and 15- and 18-membered hydrocarbon rings (MM3 and AM1 global minima are shown in Figures 2 and 3[11]). To avoid multiple duplicates in the minimized structures of these relatively small molecules, the size of the sample sets was limited to 200 structures.

† Part of the "Sason S. Shaik Festschrift".

* Authors to whom correspondence should be addressed. E-mail: noham.weinberg@ufv.ca (N.W.) and swolfe@sfu.ca (S.W.).

‡ Simon Fraser University.

§ University of the Fraser Valley.

10.1021/jp805417k CCC: $40.75
Published on Web 11/21/2008

TABLE 1: Performance of the Uniform and Biased Samplings in Identifying Low-Energy MM3 Minima of 200-Molecule Samples of Hydrocarbons

	bias					
	none (uniform search)		1D ("weighted search")		2D ("correlated search")	
hydrocarbon	0−1 kcal	0−3 kcal	0−1 kcal	0−3 kcal	0−1 kcal	0−3 kcal
n-C_9H_{20}	3	28	4[a]	33	4[a]	41
n-$C_{12}H_{26}$	0	4	6[a]	113	6[a]	112
n-$C_{15}H_{32}$	0	0	7[a]	131	7[a]	126
n-$C_{18}H_{38}$	0	0	8[b]	151	8[b]	130
cyclo-$C_{18}H_{36}$	1	8	2	18	7[c]	38

[a] All known minima. [b] Of the 9 known minima; the finding that one minimum was not found within the first 200-molecule sample reflects the increasingly large size of the conformational space of a flexible chain molecule. [c] The actual number of minima within 1 kcal range from the global minimum of cyclooctadecane is not known; for cycloheptadecane ($C_{17}H_{34}$) it is three.[4]

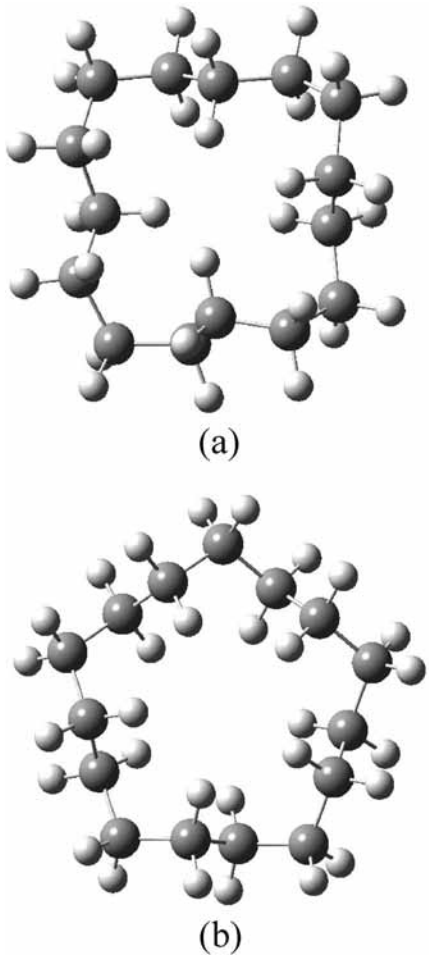

Figure 2. MM3 (a) and AM1 (b) global minima of cyclopentadecane. Interestingly, AM1 shows preference for a minimum of higher symmetry.

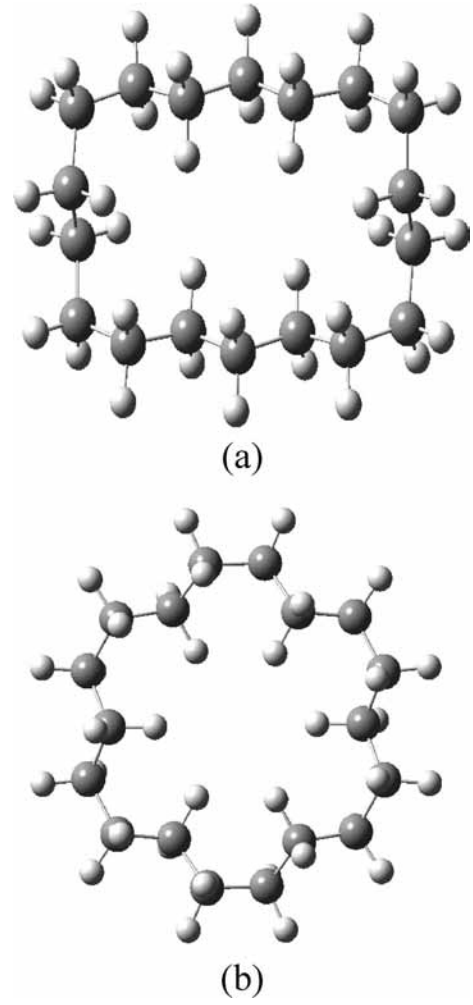

Figure 3. MM3 (a) and AM1 (b) global minima of cyclooctadecane.[11]

3. The Method

The objective is straightforward—to sample efficiently the entire conformational space or its essential parts by creating a bias favoring structures of interest. This is achieved by generating sample conformations with torsion angles randomly selected according to a specified probability distribution function. This biasing distribution function can be obtained either from the experimental structural data or through the frequency analysis of dihedral angles in a uniformly generated sample set. If necessary, this estimate can be further improved self-consistently in a sequence of iterations involving biased samples (see the flow chart of Figure 4). We used one- (1D) and two-dimensional (2D) torsion angle distribution functions. The latter, well-known in protein conformation analysis as Ramachandran maps,[12] allow for correlations between neighboring torsion angles. In the 1D case the random torsion angles were generated independently of each other. This difference did not seem to be important for open chains of relatively small size, but had a notable effect in cyclic structures (Table 1).

The method can be illustrated by its application to the search of low-energy conformations of *n*-pentadecane. First, a uniform sample of 200 conformations was generated and minimized, producing a rather broad energy distribution centered at about 8 kcal/mol and ranging from 3 to 12 kcal/mol (Figure 5). The first-generation bias (Figure 6) was then extracted from the structural information on the lowest 3-kcal minima of this

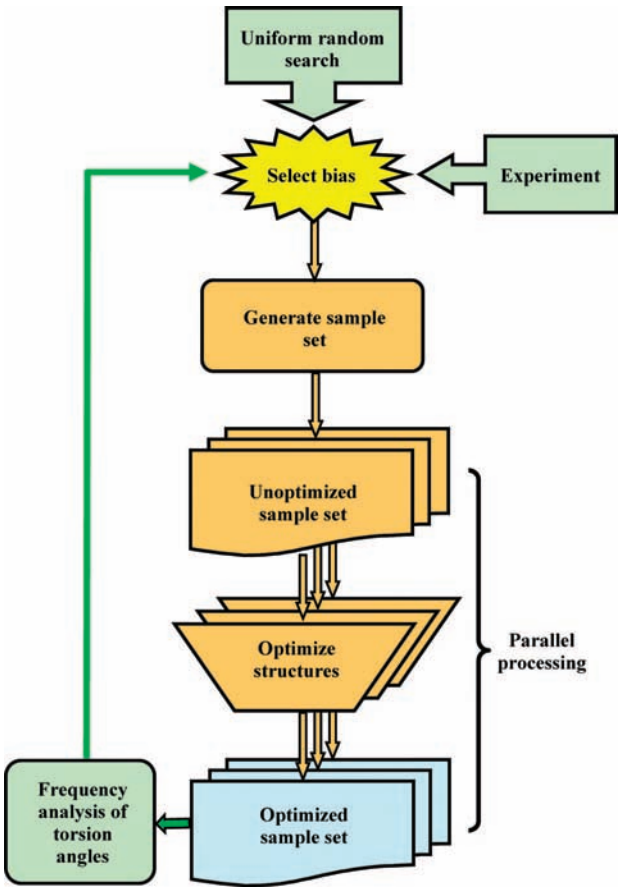

Figure 4. Energy distributions for 200 MM3-minimized random conformations of *n*-pentadecane (uniform and biased samples).

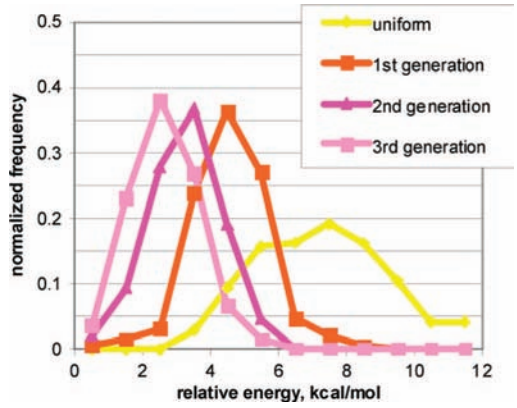

Figure 5. A flow chart of the importance sampling conformational search. The torsion angle probability distribution bias is obtained either from the experimental data or from a preceding cycle of random conformational search (uniform or biased).

uniformly sampled set and used to produce a first-generation biased sample of 200 conformations, which was also minimized, producing a much narrower energy distribution centered at about 5 kcal/mol (Figure 5). A refined (second-generation) bias was then extracted from the lowest 3-kcal minima of the biased set of the first generation, and so on. The biasing probability distribution functions were produced over a discrete 10° torsional angle grid:

$$P(\tau) = \frac{N(\tau - 5^\circ; \tau + 5^\circ)}{N_{\text{total}}}; \quad \tau = 10^\circ n \quad (n = -18, ..., +18)$$

For example, the probability of a 60° torsion $P(60°)$ was estimated as the fraction of dihedral angles greater than 55°

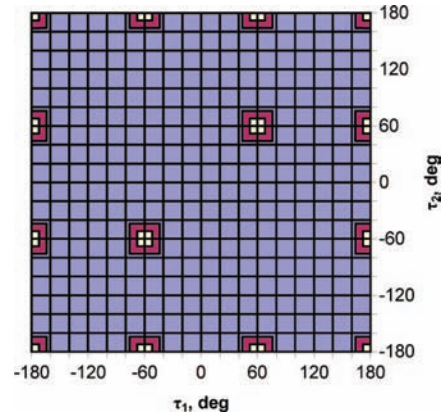

Figure 6. First-generation bias: A Ramachandran-like 2D distribution function for the neighboring dihedral angles τ_1 and τ_2 extracted from the lowest 3-kcal subset of 200 MM3-minimized uniformly sampled random conformations of *n*-pentadecane.

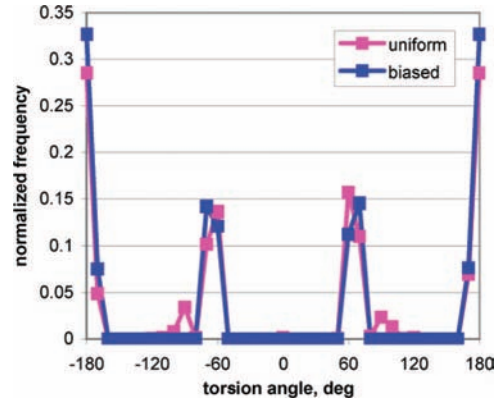

Figure 7. 1D distribution functions for the dihedral angles in 200 MM3-minimized random conformations of *n*-pentadecane (uniform and biased samples).

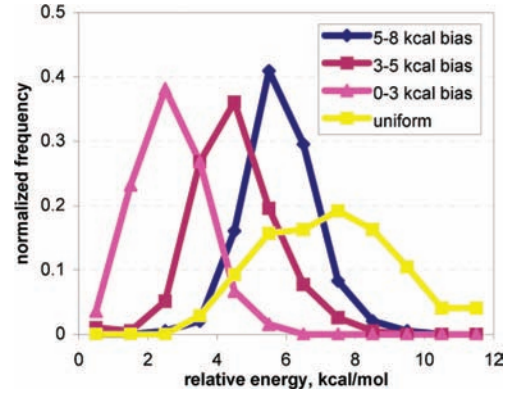

Figure 8. Energy distributions for 200 MM3-minimized random conformations of *n*-pentadecane (uniform sample and samples biased toward particular energy ranges).

but less than or equal to 65°:

$$P(60^\circ) = N(55^\circ < \tau \leq 65^\circ)/N_{\text{total}}.$$

The dihedral angle distributions for the initial uniform and the final biased sample are fairly close, although the former displays semieclipsed conformations with ± 90° torsions in 1D distribution (Figure 7), and energy-unfavorable g^+g^- junctions in 2D distribution.

The energy distributions show a remarkable sensitivity to the choice of the biasing function. As seen in Figure 8, a biasing

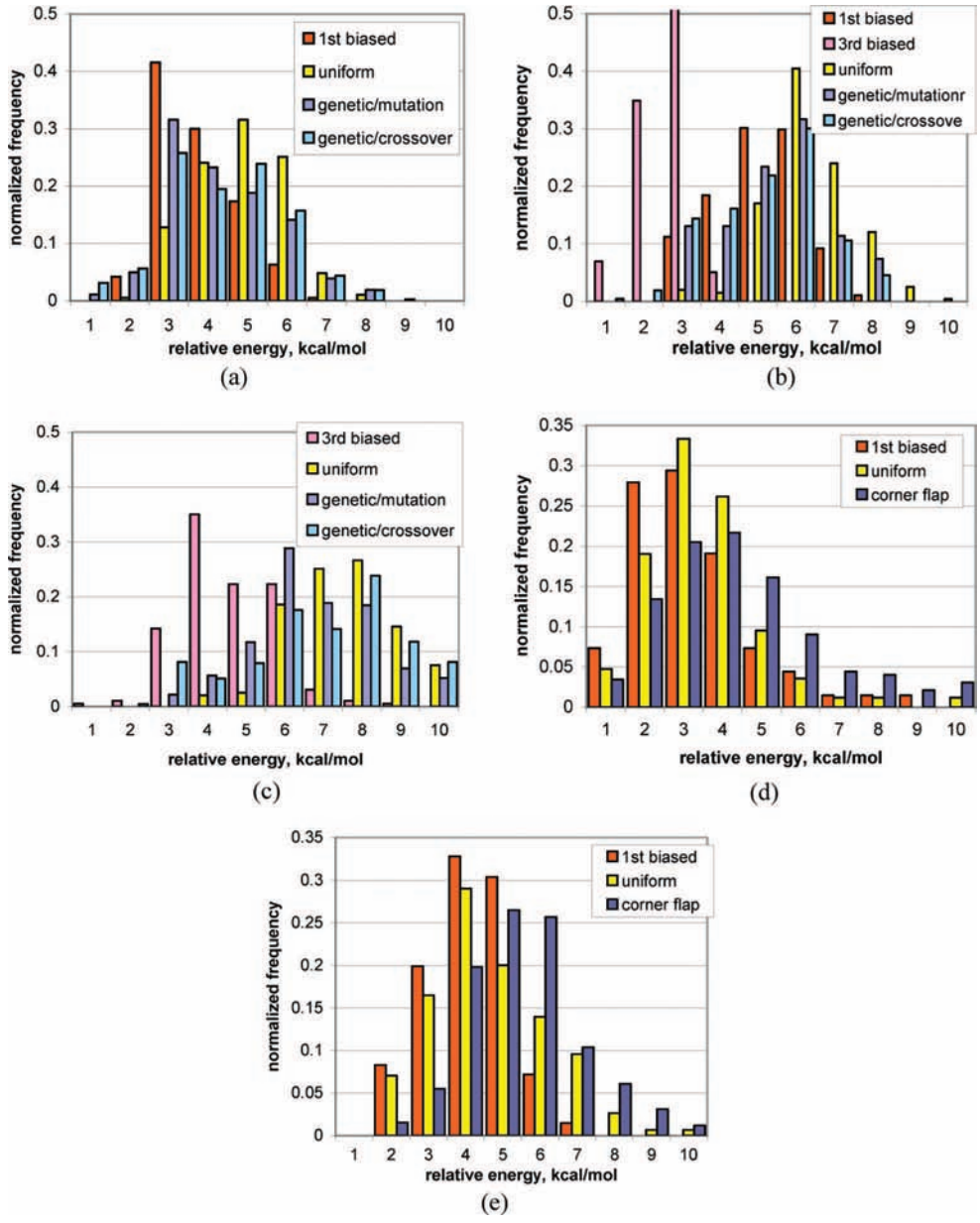

Figure 9. Energy distributions for 200 AM1-minimized random conformations of hydrocarbons obtained by different techniques: (a) *n*-dodecane, (b) *n*-pentadecane, (c) *n*-octadecane, (d) cyclopentadecane, and (e) cyclooctadecane

function extracted from the conformations of a given energy range produces a narrow energy distribution centered within that range.

4. Performance of the Method

The performance of the method was tested in a search of low-energy conformations of hydrocarbon chains and rings in comparison with the uniform random search,[4] the corner flap method,[9] and two earlier mentioned simplified genetic algorithms[2] with operations restricted to mutations only or crossovers only. The results are shown in Figure 9.

In all cases, importance sampling, even with the first-generation bias, produced distributions of minima shifted further toward low-energy conformations than other techniques. Neither method was able to locate the global minima of *n*-pentadecane, *n*-octadecane, and cyclooctadecane in the first batch of 200 structures. However, all of these minima were present in the third-generation biased sample obtained as described in the previous section.

5. General Strategy

Equipped with a proper biasing function, importance sampling offers a significant improvement over a uniform random search, while maintaining global coverage of the conformational space or its parts and suitability for parallel computing. The first-guess biasing function can be obtained from the available experimental data, preliminary uniform random search, or a combination of both. If necessary, this function can be refined iteratively in a sequence of biased searches. The most promising leads revealed by the importance sampling can be pursued further by using random walk techniques or genetic algorithms. Such combinations of search techniques would provide a fast, highly parallelizable, and focused method for the conformational analysis of flexible systems. For complex molecules including different structural elements, such as helices, strands, or loops, the biasing function can be chosen differently for different segments of the molecule to reflect its structural inhomogeneity.

Acknowledgment. The authors thank NSERC, the Natural Sciences and Engineering Research Council of Canada, for financial support of this research and WestGrid, the Western Canada Research Grid, for allocation of computational resources.

References and Notes

(1) Leach, A. *Molecular Modelling: Principles and Applications*; Prentice Hall: Upper Saddle River, NJ, 2001.

(2) (a) Nair, N.; Goodman, J. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317, and references cited therein. See also: (b) Holland, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*; University of Michigan Press: Ann Arbor, MI, 1975. (c) Mitchell, M. *An Introduction to Genetic Algorithms*; The MIT Press: Cambridge, MA, 1996.

(3) (a) Paine, G. H.; Scheraga, H. A. *Biopolymers* **1985**, *24*, 1391. (b) Paine, G. H.; Scheraga, H. A. *Biopolymers* **1986**, *25*, 1547.

(4) Weinberg, N.; Wolfe, S. *J. Am. Chem. Soc.* **1994**, *116*, 9860, and references cited therein.

(5) (a) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087. (b) Hastings, W. K. *Biometrika* **1970**, *57*, 97. (c) Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*; Chapman & Hall: London, UK, 1997. (d) Landau, D. P.; Binder, K. *A Guide to Monte Carlo Simulations in Statistical Physics*; Cambridge University Press: Cambridge, UK, 2005. (e) *Free Energy Calculations. Theory and Applications in Chemistry and Biology*; Chipot, Ch., Pohorille, A., Eds.; Springer-Verlag: Berlin, Germany, 2007.

(6) *Operating Instructions for MM3 Program*; Technical Utilization Corporation: 235 Glen Village Court, Powell, OH, 1989. See: Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(7) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *I07*, 3902.

(8) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(9) Hitoshi Goto, H.; Osawa, E. *J. Am. Chem. Soc.* **1989**, *I I I*, 8950.

(10) Goodman, J. M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 876.

(11) See also MM2 results for cyclooctadecane in: Shah, A. V.; Dolata, D. P. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 103.

(12) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963**, *7*, 95.

JP805417K