Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

LIONEL WALTER, MARCUS WEBER

ConfJump : a fast biomolecular sampling method which drills tunnels through high mountains

ZIB-Report 06-26 (June 2006)

ConfJump : a fast biomolecular sampling method which drills tunnels through high mountains

Lionel Walter and Marcus Weber

June 22, 2006

Abstract

In order to compute the thermodynamic weights of the different metastable conformations of a molecule, we want to approximate the molecule's Boltzmann distribution π in a reasonable time. This is an essential issue in computational drug design. The energy landscape of active biomolecules is generally very rough with a lot of high barriers and low regions. Many of the algorithms that perform such samplings (e.g. the hybrid Monte Carlo method) have difficulties with such landscapes. They are trapped in low-energy regions for a very long time and cannot overcome high barriers. Moving from one low-energy region to another is a very rare event. For these reasons, the distribution of the generated sampling points converges very slowly against the thermodynamically correct distribution of the molecule.

The idea of ConfJump is to use *a priori* knowledge of the localization of low-energy regions to enhance the sampling with artificial jumps between these low-energy regions. The artificial jumps are combined with the hybrid Monte Carlo method. This allows the computation of some dynamical properties of the molecule. In ConfJump, the detailed balance condition is satisfied and the mathematically correct molecular distribution is sampled.

MSC: 65C05, 60J22, 92E10, 37A60
Keywords: Monte Carlo simulation, rare events, rough potential energy function, molecular dynamics

1 Introduction

As a research group for computational drug design, we examine the binding capacity of different ligands to certain target molecules [7, 8]. Since the behavior of a single ligand molecule depends on its structure, we want to determine those conformations which are suitable for the docking process [2, 3, 10]. Therefore, if $\Omega = \mathbb{R}^{3n}$ is the configuration space for a molecule of n atoms, the first step consists in the identification of all metastable conformations $C_1, \ldots C_m \subset \Omega$ as defined in [20]. We also want to know their thermodynamic weights $w_1, \ldots, w_m$. These are the probabilities for the molecule to be in a given conformation, i.e.

$$w_i = \frac{\int_{C_i} \pi(q)dq}{\int_\Omega \pi(q)dq}, \quad i = 1, \ldots, m.$$

To this aim we need to sample from the Boltzmann distribution $\pi(q)$ of the molecule. Furthermore, to identify the metastable conformations we need dynamic information, i.e. knowledge about the molecule's behavior over time.

Conformation dynamics based on hybrid Monte Carlo techniques [20] and the transfer operator approach [21] are powerful methods for conformation analysis. The software package ZIBgridfree [15, 17, 25, 26], based upon classes of amiraMol [19], combines these ideas with meshfree methods. It does not only determine metastable conformations and their thermodynamic weights, but also the correct transition probabilities. However, this algorithm is very expensive because it is based on a correct sampling of the Boltzmann distribution even in very sparsely occupied transition regions. This problem, called "broken ergodicity" or "critical slowing down", arises when one wants to sample from a very rough distribution such as the ones in molecular dynamics. These distributions are related to the potential energy of the molecule as follows :

$$\pi(q) = \frac{\exp(-\beta V(q))}{\int_\Omega \exp(-\beta V(q)) dq},$$

where $q \in \Omega$ is a molecular configuration for a molecule of n atoms, $\Omega = \mathbb{R}^{3n}$ the configuration space, $\beta = \frac{1}{k_B T}$ the inverse temperature and $V(q)$ the potential energy. Such potential energies functions have typically very high barriers between the low-energy regions. Thus transitions from one low-energy region to another are rare events. Often simulations are trapped in low-energy regions and very long simulations are necessary to sample the whole configuration space.

In ConfJump, we search at the beginning for all regions of low energy. Then we use this information to enhance the sampling algorithm, as if we were drilling tunnels through the high energy barriers. All that is done in a mathematical rigorous way to ensure the detailed balance condition.

Various so-called *smart* methods have been designed to overcome the "broken ergodicity" problem, see [4] for a survey. The methods which are closely related to our method are the Smart Darting Monte Carlo [1] and the Jump Between Wells (JBW) method [22, 23]. Our method can be seen as an extension of the latter method, with two main differences :

- The jump steps are combined with the hybrid Monte Carlo method. So we not only have a sampling but dynamic information as well. This allows e.g. the computation of metastable conformations.
- The use of a jump proposition matrix (cf. 2.2) considerably enhances the acceptance rate for the jump steps.

In the next subsection, we explain our motivation for developing ConfJump . In section 2 we describe the algorithm and in section 3 we give results from our computer simulations.

1.1 Higher-order Integrators

In order to obtain a better convergence in ZIBgridfree, we wanted to improve the acceptance of the hybrid Monte Carlo steps. Therefore, we implemented symplectic integrators of higher order than the velocity Verlet integrator which was already implemented. We tested the composition integrator described in [11]. Two problems arose :

- The better acceptance didn't imply a better convergence. The "mistakes" of the velocity Verlet integrator are somehow good for moving towards other regions of low-energy.
- To use an integrator of higher order with the same cost as the velocity Verlet integrator, we need to use longer timesteps. Unfortunately, stability problems often result from the use of long timesteps and that was the case in our molecular dynamics simulations.

The above reasons led us to ConfJump, a completely different approach to improve the convergence of our simulations.

2 The ConfJump algorithm

A description of the ConfJump algorithm can be found in figure 1. It is a Monte Carlo method which combines 2 kinds of steps :

- hybrid Monte Carlo steps to ensure ergodicity and compute metastable regions and
- jump steps to accelerate convergence.

The following subsections explain details about the algorithm.

Preprocessing :

1. Perform a conformational search to find a set $M = (m_1, m_2, \ldots, m_N)$ of N low-energy configurations. The configurations are given in internal coordinates (Z matrix) (cf. 2.1).
2. Compute jump proposition matrix A (cf. 2.2).
3. Select an initial configuration $q \in \Omega$ (in Cartesian coordinates).

Until Convergence (Gelman-Rubin Criterion), repeat :

1. Pick a random number z in $(0, 1)$.
2. If $z >$ jumpRate, perform a hybrid Monte Carlo step.
3. Else, perform a jump step :
   (a) Find[1]the nearest low-energy configuration $m_i = f(q)$ in the list M.
   (b) Select another low-energy configuration m_j with probability A_{ij}.
   (c) Compute x, the internal coordinates (Z matrix) of q.
   (d) Compute $\tilde{x} = x + (m_j - m_i)$, and transform $\tilde{x}$ to Cartesian coordinates to get the new point $\tilde{q}$.
   (e) Check that $f(\tilde{q}) = m_j$, i.e. that the nearest low-energy configuration of the new point is the configuration selected in 3b. If not reject the point (i.e. set $\tilde{q} = q$).
   (f) Accept the new point with probability $\min\left(1, \mathrm{e}^{-\beta(V(\tilde{q})-V(q))}\right)$ (Metropolis acceptance criterion).

Postprocessing :

1. Identify the metastable conformations using successive PCCA [6] based only on the hybrid Monte Carlo steps.
2. Compute weights $\left(w_i \approx \dfrac{\#\text{ points in conformation } i}{\text{total }\#\text{ points}}\right)$.

Figure 1: The ConfJump algorithm

2.1 Search for low-energy regions

The first step of ConfJump is the identification of all the low-energy regions. This is a challenging task and a lot of methods have been developed with this aim [5]. Such methods search for the numerous minima of the potential energy. They give representatives of the low-energy regions. But those methods cannot compute the thermodynamic weights of the different conformations.

Our method of choice for finding the low-energy regions is ConFlow, which has been proven empirically to give better results than other algorithms [16]. This method supplies a set of configurations $m_1, m_2, \ldots, m_N \in \Omega$ which are representatives of all the low-energy regions of a molecule. To get a good jump acceptance rate, it is important that these points have a very low-energy, i.e. that they are minima of the potential energy function V.

[1]We consider only the flexible torsion angles which describe the conformational space. The nearest minimum $f(q)$ of a configuration $q \in \Omega$ is the one which has the smallest cyclic Euclidian distance for these given torsion angles [15].

2.2 Jump proposition matrix A

A_{ij} is the probability to jump from the region of m_i to the region of m_j. As the acceptance rate (step 3f) is related to

$$\mathrm{e}^{-\beta(V(\tilde{q})-V(q))},$$

it is better to jump towards a region of similar or lower energy to get a good acceptance rate. A must be stochastic because

$$1=\sum_{j=1}^{N} P(\text{jump from } m_i \text{ to } m_j)=\sum_{j=1}^{N} A_{ij}.$$

In addition, to ensure the detailed balance condition for Metropolis Monte Carlo (symmetric proposition criterion), we need that

$$P(\text{jump from } m_i \text{ to } m_j)=P(\text{jump from } m_j \text{ to } m_i)$$

i.e. that A is symmetric. Therefore A must be a double stochastic matrix. This is the only condition on A. As long as A is a double stochastic matrix, the values of A are entirely free. An idea could be to favor jumps towards regions of lower energy. Assume m_j has a lower energy than m_i. It is good when A_{ij} is high because jumps towards regions of lower energy are often accepted. As A must be symmetric, A_{ji} must be high too. That means that the jumps towards regions of higher energy will be favored too. We don't want this effect because these jumps are almost always rejected. Therefore we decide to favor jumps between regions of similar energy. Another advantage is that regions of similar energy often have similar "shapes" (cf. 3 and figure 4). Thus, jumps between these regions often succeed.

To implement this behavior, we set

$$\begin{aligned}\hat{A}_{ij} &= \mathrm{e}^{-\beta|V(m_j)-V(m_i)|} \quad 1\le i,j\le N,\ i\ne j\\ \hat{A}_{ii} &= 0 \quad 1\le i\le N \text{ (we never jump towards the current region)}\end{aligned}$$

Then we use Ruiz's algorithm [18] to scale the matrix $\hat{A}$ to obtain the jump proposition matrix A. If $\hat{A}$ is nondegenerate, i.e. $\hat{A}_{ij}\ne 0\ \forall i\ne j$, this is equivalent to finding the unique diagonal matrix D s.t. $A=D\hat{A}D$ is double stochastic [14]. Note that within this process all zero entries remain zero entries, in particular $A_{ii}=0$, $\forall i=1,\dots,N$.

In the hybrid Monte Carlo steps, all kinds of jumps occur as well. There are jumps towards regions of lower, higher or similar energy. But all these jumps are rare events. With the jump steps, we do these jumps artificially and we can decide which kind of jumps we favor.

Cyclohexane is a good example where the jump proposition matrix is useful. Cyclohexane has 2 conformations of very low energy (called chair conformations) and 6 conformations of higher energy (called boat). When one uses only standard hybrid Monte Carlo, the probabilities to go from a conformation to another are described in figure 2 left. As the transitions between conformations are rare events, they arise not often. When one uses ConfJump, one gets the results from figure 2 right.

The transitions between conformations of equal energy arise much more often. The other transitions don't change a lot. That means that transitions between boat and chair happen mainly in the hybrid Monte Carlo steps. In table 1, we compare the random strategy, where we jump to a randomly chosen region (other than the same region), and the strategy that employs the jump proposition matrix.

Jump Acceptance Rates (in %)

Given points	random	with jump proposition matrix
2 chairs	3.9	3.9
2 chairs + 6 boats	0.8	5.3

Table 1: Comparing the 2 different strategies to choose the new region

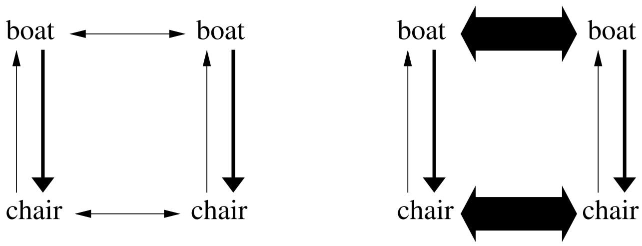

Figure 2: Probabilities to move from a conformation to another. On the left using only hybrid Monte Carlo, on the right using ConfJump. The thickness of the arrows indicates the probability to go from a conformation to another

When the high-energy boat conformations are given as input, the jump acceptance rate is more than 5 times bigger using the jump proposition matrix. The observed behavior has also been verified on bigger molecules. The average jump acceptance rate for the 32 molecules (cf. 3) that we studied increases from 9% to 19% when we use the jump proposition matrix.

2.3 Detailed Balance Condition

ConfJump is ergodic through its hybrid Monte Carlo part. Since the jump proposition matrix A is symmetric, the jump step satisfies the detailed balance condition (note the importance of step 3e of the algorithm for this aim). The hybrid Monte Carlo method satisfies the detailed balance condition. Therefore ConfJump converges to the correct distribution π.

3 Results

We want to compare the ConfJump algorithm with the hybrid Monte Carlo method. All the simulations are conducted with a jumpRate of 20% at a temperature of 300K. For the hybrid Monte Carlo steps, we use 60 molecular dynamics steps of 1.3fs.

First we did a test for 2 simple molecules where the exact thermodynamic weights are known: butane and cyclohexane. For butane we did simulations of 5,000 steps and for cyclohexane 50,000 steps.

	butane			cyclohexane	
conformations	gauche$^-$	trans	gauche$^+$	chair 1	chair 2
exact[2]	16	68	16	50	50
hybrid Monte Carlo	27	53	20	62	38
ConfJump	19	64	16	52	48

Table 2: Comparison of the thermodynamic weights (in %).

Cyclohexane is a really good example for the efficiency of ConfJump because the barrier between the 2 chair conformations is very high. Therefore the moves in hybrid Monte Carlo from one conformation to another are very rare.

We go on by testing ConfJump on the 32 molecules listed in [5]. This is a typical set of drug-like molecules, which have between one and eleven freely rotatable bonds.

We did a simulation with 20'000 steps of ConfJump. Then we ran exactly the same program with a jumpRate of zero, thus performing only standard hybrid Monte Carlo. We ran this simulation for the same amount of CPU time (on a 1,6 MHz processor) needed by ConfJump (including minima search). We did the Monte-Carlo simulation using 5 Markov chains with different starting

[2] from a very long simulation

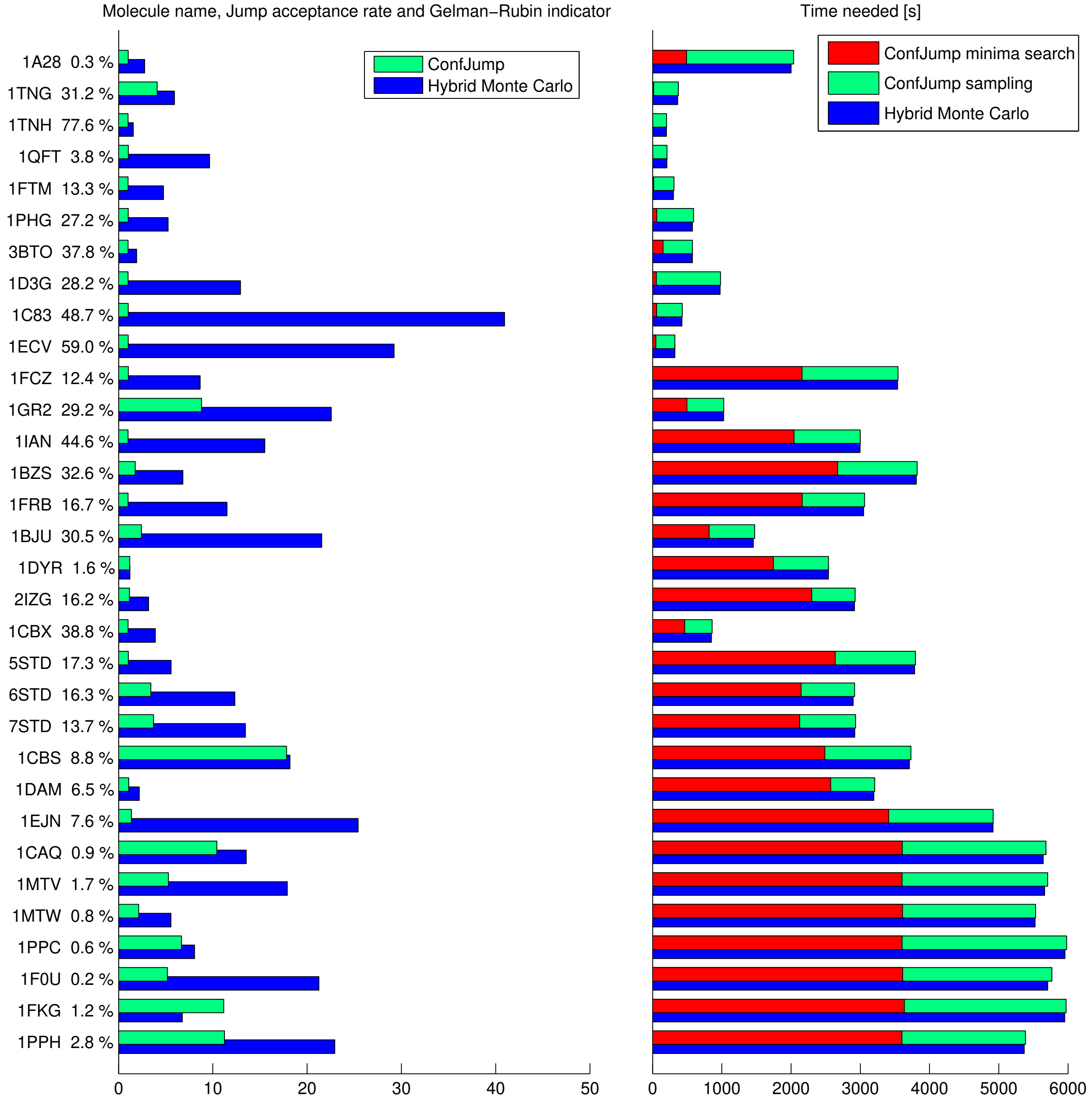

Figure 3: Comparison of ConfJump and hybrid Monte Carlo. Molecules sorted by increasing number of flexible (rotatable) torsion angles

configurations. Thus, we can compute the Gelman-Rubin convergence indicator [9]. The Gelman-Rubin convergence criterion compares the total variance of the 5 chains with the variance of each individual chain. Gelman-Rubin convergence indicator close to one means convergence. The results are shown in figure 3.

They clearly show ConfJump's superiority over the standard hybrid Monte Carlo approach. For all molecules but one (1FKG) ConfJump was better than the hybrid Monte Carlo method. For a given amount of time, it is much better to perform first a minima search and then a sampling using the results of that search than conducting only a sampling in the hope that the algorithm will find all regions of low-energy on its own.

The rate of accepted jumps varies between 0.2% and 77.5%. It decreases while the number of flexible torsion angles increases. The geometric similarity of conformations is important for the jump acceptance rate. When the conformations have more or less the same "shape" in internal coordinates, than the jump acceptance rate is better (see figure 4). Currently, we try to gather information about the "shapes" of the conformations during the minima search phase. We then want to use this information in the jump steps, e.g. as described in [24].

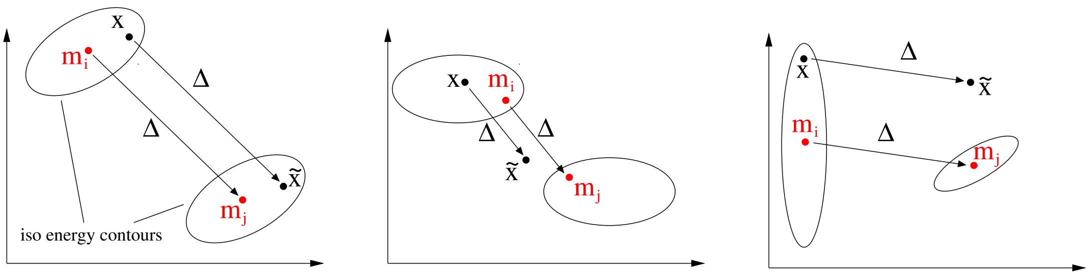

Figure 4: Three different situations for a jump step. The two axes represent two given internal coordinates. On the left a jump which succeeds. The two conformations have more or less the same shape. In the middle the jump fails because the relative positions of m_i and m_j in the conformation are different. On the right the jump fails because the conformations have very different shapes.

4 Conclusion

We have shown that performing a minima search and then using the information obtained that way in the sampling phase is much more efficient than performing only a sampling. The advantage of using the hybrid Monte Carlo method combined with jumps is that at the end we can still compute dynamical properties such as metastable conformations and transition probabilities. We just do these analyses on all the hybrid Monte Carlo steps ignoring the steps with jumps. The major importance of the jump proposition matrix to enhance the acceptance rate must be emphasized as well. It is worth to note that ConfJump can easily be generalized to other domains where the "critical slowing down" problem arises.

Acknowledgments. The authors want to thank Peter Deuflhard for the always inspiring discussions about higher-order integrators and conformation dynamics, Frank Cordes for his help in the development of ConfJump, Holger Meyer, the author of ConFlow, whose very good minima are of major importance in the success of ConfJump, Alexander Riemer for his explanations about algorithmic details of ZIBgridfree, Johannes Schmidt-Ehrenberg for the adaption of the successive PCCA algorithm to ignore jump steps and for his patience in explaining details of amiraMol and Daniel Baum for his deep knowledge of amiraMol.

References

[1] I. Andricioaei, J. Straub, and A. Voter. Smart darting monte carlo. *Journal of Chemical Physics*, 114(16):6994–7000, 2001.

[2] D. Baum. Multiple semi-flexible 3d superposition of drug-sized molecules. Technical Report 04-52, Zuse Institute Berlin (ZIB), 2004.

[3] T. Baumeister and F. Cordes. A new model for the free energy of solvation and its application in protein ligand scoring. Technical Report 04-51, Zuse Institute Berlin (ZIB), 2004.

[4] B. Berne and J. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Current Opinion in Structural Biology*, 7:181–189, 1997.

[5] J. Boström. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *Journal of Computer-Aided Molecular Design*, 15(12):1137–1152, 2001.

[6] F. Cordes, M. Weber, and J. Schmidt-Ehrenberg. Metastable conformations via successive Perron-cluster cluster analysis of dihedrals. Technical Report ZIB 02-40, Zuse Institute Berlin, 2002.

[7] P. Deuflhard. From molecular dynamics to conformational dynamics in drug design. In M. Kirkilionis, S. Krömker, R. Rannacher, and F. Tomi, editors, *Trends in Nonlinear Analysis*, pages 269–287. Springer, 2003.

[8] P. Deuflhard and Ch. Schütte. Molecular Conformation Dynamics and Computational Drug Design. In J. Hill and R. More, editors, *Applied Mathematics Entering the 21st Century, Invited Talks from the ICIAM 2003 Congress*, pages 91–119. 2004.

[9] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Sience*, 7:457–511, 1992.

[10] A. Gürler. Selection and Flexible Optimization of Binding Modes from Conformation Ensembles. Master's thesis, Free University Berlin, April 2006.

[11] E. Hairer, Ch. Lubich, and G. Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations.* Springer Series in Computational Mathematics. 31. Berlin: Springer., 2002.

[12] T.A. Halgren. *J. Am. Chem. Soc.*, 114:7827–7843, 1992.

[13] T.A. Halgren. Merck molecular force field. *J. Comp. Chem.*, 17(I-V):490–641, 1996.

[14] C. Johnson, R. Masson, and M. Trosset. On the diagonal scaling of Euclidean distance matrices to doubly stochastic matrices. *Linear Algebra Appl.*, 397:253–264, 2005.

[15] H. Meyer. Die Implementierung und Analyse von HuMfree–einer gitterfreien Methode zur Konformationsanalyse von Wirkstoffmolekülen. Master's thesis, Free University Berlin, February 2005.

[16] H. Meyer., F. Cordes, and M. Weber. ConFlow – a new space-based application for complete conformational analysis of molecules. Technical Report 06-31, Zuse Institute Berlin, 2006.

[17] H. Meyer, M. Weber, and A. Riemer. Zibgridfree. Software package for HMC-simulation and conformation analysis based upon C++ classes of amiraMol [19] using the Merck Molecular Force Field [12, 13] implemented by T. Baumeister and parametrized by F. Cordes. Robust Perron Cluster Analysis implemented by M. Weber and J. Schmidt-Ehrenberg., Status: January 2006. Software owned by the Zuse Institute Berlin.

[18] D. Ruiz. A Scaling Algorithm to Equilibrate both Rows and Columns Norms in Matrices. Technical Report Technical Report RAL-TR-2001-034, Rutherford Appleton Laboratory, 2001.

[19] J. Schmidt-Ehrenberg, D. Baum, and H.-Ch. Hege. Visualizing dynamic molecular conformations. In *IEEE Visualization 2002*, pages 235–242. IEEE Computer Society Press, 2002.

[20] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys., Special Issue on Computational Biophysics*, 151:146–168, 1999.

[21] Ch. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In B. Fiedler, editor, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems.*, pages 191–223. Springer, 2001. Preprint available via http://www.math.fu-berlin.de/˜biocomp.

[22] H. Senderowitz, F. Guarnieri, and W. C. Still. A Smart Monte Carlo Technique for Free Energy Simulations of Multiconformational Molecules. Direct Calculations of the Conformational Populations of Organic Molecules. *Journal of the American Chemical Society*, (117):8211–8219, 1995.

[23] H. Senderowitz and W. C. Still. MC(JBW): Simple but smart monte carlo algorithm for free energy simulations of multiconformational molecules. *Journal of Computational Chemistry*, 19(15):1736–1745, 1998.

[24] C. Sminchisescu, M. Welling, and G. Hinton. A Mode-Hopping MCMC Sampler. Technical Report CSRG-478, University of Toronto, 2003.

[25] M. Weber. *Meshless Methods in Conformation Dynamics*. PhD thesis, Free University Berlin, department of mathematics, 2006.

[26] M. Weber and H. Meyer. ZIBgridfree - adaptive conformation analysis with qualified support of transition states and thermodynamic weights. Technical Report ZIB 05-17, Zuse Institute Berlin, 2005.