To cite this document : Devaurs, Didier and Vaisset, Marc and Siméon, Thierry and Cortés, Juan *A multi-tree approach to compute transition paths on energy landscapes.* (In Press: 2013) In: Workshop on Artificial Intelligence and Robotics Methods in Computational Biology, AAAI '13, July, 2013, Bellevue, Washington, USA.

A Multi-Tree Approach to Compute Transition Paths on Energy Landscapes

Didier Devaurs and **Marc Vaisset** and **Thierry Siméon** and **Juan Cortés**
CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France
and Univ de Toulouse, LAAS, F-31400 Toulouse, France
{devaurs, mvaisset, nic, jcortes}@laas.fr

Abstract

Exploring the conformational energy landscape of a molecule is an important but challenging problem because of the inherent complexity of this landscape. As part of this theme, various methods have been developed to compute transition paths between stable states of a molecule. Besides the methods classically used in biophysics/biochemistry, a recent approach originating from the robotics community has proven to be an efficient tool for conformational exploration. This approach, called the Transition-based RRT (T-RRT) is based on the combination of an effective path planning algorithm (RRT) with a Monte-Carlo-like transition test. In this paper, we propose an extension to T-RRT based on a multi-tree approach, which we call *Multi-T-RRT*. It builds several trees rooted at different interesting points of the energy landscape and allows to quickly gain knowledge about possible conformational transition paths. We demonstrate this on the alanine dipeptide.

Introduction

Global thermodynamic and kinetic properties of molecules can be extracted from an analysis of their conformational energy landscape (Wales 2003). Thus, obtaining an accurate representation of this landscape is an important problem that has sparked the interest of the scientific community for decades. This problem is challenging because, in general, the energy landscape is a high-dimensional, rugged manifold. Among the issues this problem raises, two particularly interesting ones are: 1) how to achieve a significant and efficient sampling of the conformational space, and 2) how to compute transition paths between stable states of a molecule. Different approaches have been developed to explore and to represent energy landscapes, but there is still room for the development of more efficient and/or more accurate methods.

Recent work shows that algorithms originating from the field of robotics can be a good basis for efficient conformational-sampling and exploration methods in computational structural biology (Gipson et al. 2012; Al-Bluwi, Siméon, and Cortés 2012). The Transition-based RRT (T-RRT) algorithm is an example of such algorithms (Jaillet et al. 2011). It can be applied to identify interesting points on the energy landscape (i.e. minima and saddle-points) and to compute probable conformational transition paths. T-RRT is based on the Rapidly-exploring Random Tree (RRT) algorithm (LaValle and Kuffner 2001), a popular path planning algorithm that can tackle complex problems in high-dimensional spaces. RRT has been successfully used in various disciplines, such as robotics, manufacturing, computer animation and computational structural biology. T-RRT is an extension of RRT involving a probabilistic transition test based on the Metropolis criterion. In the same way as Metropolis Monte Carlo methods (Frenkel and Smit 2001), it applies small moves to the molecular system; but, instead of generating a single path over the conformational space, it constructs a tree, providing a more efficient exploration. Moreover, the tree construction is intrinsically biased toward unexplored regions of the space, and favors expansions on low-energy areas.

In this paper, we propose an extension of T-RRT (which is a single-tree algorithm) based on a multi-tree approach, which we name *Multi-T-RRT*. Instead of growing a single tree rooted at a given molecular conformation, the Multi-T-RRT grows several trees rooted at different interesting conformations (e.g. local minima) scattered over the exploration space, and provided as input. These initial conformations may correspond to experimentally-determined structures, or may be generated by computational methods, such as simulated annealing (Wilson et al. 1991), basin hopping (Wales and Doye 1997), or recent methods based on evolutionary search (Olson and Shehu 2012). The Multi-T-RRT algorithm aims at quickly providing information about regions of the energy landscape through which transition paths might be observed between the given set of conformations. To illustrate this, and as a proof-of-concept, we present initial results obtained on the alanine dipeptide, a frequent benchmark for theoretical studies in biophysics/biochemistry.

Methods

Rapidly-exploring Random Tree (RRT)

Starting from an initial conformation, q_{init}, RRT iteratively constructs a tree $\mathcal{T}$ that tends to rapidly expand on the

Algorithm 1: Transition-based RRT

```
input  : the conformational space $\mathcal{C}$
         the energy function $E : \mathcal{C} \rightarrow \mathbb{R}$
         the initial conformation $q_{init}$
         the target conformation $q_{goal}$ (optional)
         the extension step-size $\delta$
output: the tree $\mathcal{T}$
$\mathcal{T} \leftarrow$ initTree($q_{init}$)
while not stopCondition($\mathcal{T}$, $q_{goal}$) do
    $q_{rand} \leftarrow$ sampleRandomConformation($\mathcal{C}$)
    $q_{near} \leftarrow$ findNearestNeighbor($\mathcal{T}$, $q_{rand}$)
    if refinementControl($\mathcal{T}$, $q_{near}$, $q_{rand}$) then
        $q_{new} \leftarrow$ extend($q_{near}$, $q_{rand}$, $\delta$)
        if $q_{new} \neq null$ and
        transitionTest($\mathcal{T}$, $E(q_{near})$, $E(q_{new})$) then
            addNewNode($\mathcal{T}$, $q_{new}$)
            addNewEdge($\mathcal{T}$, $q_{near}$, $q_{new}$)
```

Algorithm 2: transitionTest ($\mathcal{T}$, E_i, E_j)

```
input  : the energy threshold $E_{max}$
         the current temperature $T$
         the temperature increase rate $T_{rate}$
output: true if the transition is accepted, false if not
if $E_j > E_{max}$ then return False
if $E_j \leq E_i$ then return True
if $exp(-(E_j - E_i) \,/\, (K \cdot T)) > 0.5$ then
    $T \leftarrow T \,/\, 2^{(E_j - E_i) \,/\, (0.1 \cdot \texttt{energyRange}(\mathcal{T}))}$
    return True
else
    $T \leftarrow T \cdot 2^{T_{rate}}$
    return False
```

Algorithm 3: refinementControl ($\mathcal{T}$, q_{near}, q_{rand})

```
input  : the extension step-size $\delta$
         the refinement ratio $\rho$
output: true if refinement is low enough, false if not
if distance($q_{near}$, $q_{rand}$) $< \delta$ and
nbRefinementNodes($\mathcal{T}$) $> \rho \cdot$ nbNodes($\mathcal{T}$) then
    return False
return True
```

conformational space $\mathcal{C}$, thanks to the implicit enforcement of a Voronoi bias (LaValle and Kuffner 2001). The nodes and edges of $\mathcal{T}$ correspond to states (i.e. molecular conformations) and small-amplitude moves between states, respectively. At each iteration of the tree construction, a conformation q_{rand} is randomly sampled in $\mathcal{C}$. Then, an extension toward q_{rand} is attempted, starting from its nearest neighbor, q_{near}, in $\mathcal{T}$. This means performing an interpolation between q_{near} and q_{rand}, at a distance equal to the extension step-size, δ, from q_{near} (except if distance$(q_{near}, q_{rand}) < \delta$, in which case the result of the interpolation is q_{rand} itself). If the extension succeeds, a new conformation q_{new} is added to $\mathcal{T}$ and an edge is built between q_{near} and q_{new}. The criteria on when to stop the exploration can be reaching a given target conformation q_{goal}, a given number of nodes in the tree, a given number of iterations, or a given running time.

Transition-based RRT (T-RRT)

Contrary to RRT, T-RRT allows to explore a conformational space over which an energy function is defined. T-RRT (whose pseudo-code is shown in Algorithm 1) extends RRT by integrating a stochastic transition test enabling it to steer the exploration toward low-energy regions of the conformational space (Jaillet et al. 2011). Similarly to the Metropolis criterion typically used by Monte Carlo simulations in statistical physics (Frenkel and Smit 2001), this transition test is used to accept or reject a candidate state, based on the energy variation associated with the local move from the previous state to this state. Compared with RRT, T-RRT also features a refinement-control mechanism that will be detailed in the sequel.

The transitionTest presented in Algorithm 2 is used to evaluate the transition between the conformations q_{near} and q_{new} based on their respective energies. Three cases are possible: 1) A new conformation whose energy is higher than the threshold value E_{max} is automatically rejected. 2) A transition corresponding to a downhill move ($E_j \leq E_i$) is always accepted. 3) Uphill transitions are accepted or rejected based on the probability $exp(-(E_j - E_i) \,/\, (K \cdot T))$ (where K is the Boltzmann constant), which decreases exponentially with the energy variation $E_j - E_i$. In that case, the level of difficulty of the transition test is controlled by the *temperature* T, which is an adaptive parameter of the algorithm. Low temperatures limit the expansion to gentle slopes, and high temperatures enable to climb steep slopes. The temperature is dynamically tuned during the search process, which allows T-RRT to automatically balance its tendency to steer the exploration toward low-energy regions with the Voronoi bias of RRT. After each accepted uphill transition, T is decreased to avoid over-exploring high-energy regions. More precisely, it is divided by $2^{(E_j - E_i) \,/\, (0.1 \cdot \texttt{energyRange}(\mathcal{T}))}$, where energyRange($\mathcal{T}$) is the energy difference between the highest-energy and the lowest-energy conformations in the tree. After each rejected uphill transition, T is increased to facilitate the exploration and to avoid being trapped in a local minimum. More precisely, it is multiplied by $2^{T_{rate}}$, where $T_{rate} \in \,]0, 1]$ is the temperature increase rate. The T_{rate} parameter determines a trade-off between low computation time and low energy of the produced paths: A value not too small (e.g. 0.1) leads to a greedy search, and a lower value (e.g. 0.01) enables to produce lower-energy paths. In the rest of the paper, we use only these two values for T_{rate}.

The adaptive temperature tuning of T-RRT ensures a given success rate for uphill transitions, which can also contribute to refining the exploration of low-energy regions already reached by the tree, as a side effect. The objective of the refinementControl function (shown in Algorithm 3) is to limit this refinement and facilitate the tree expansion toward unexplored regions. The idea is to reject an expansion that would lead to more refinement if the num-

Algorithm 4: Multi-T-RRT

```
input  : the conformational space $\mathcal{C}$
         the energy function $E : \mathcal{C} \rightarrow \mathbb{R}$
         the initial conformations $q^k_{init}$, $k = 1..n$
         the extension step-size $\delta$
output: the tree $\mathcal{T}$
for $k = 1..n$ do
    $\mathcal{T}_k \leftarrow$ initTree($q^k_{init}$)
while not stopCondition($\{\mathcal{T}_k \,|\, k = 1..n\}$) do
    $\mathcal{T}' \leftarrow$ chooseNextTreeToExpand()
    $q_{rand} \leftarrow$ sampleRandomConfiguration($\mathcal{C}$)
    $q'_{near} \leftarrow$ findNearestNeighbor($\mathcal{T}'$, $q_{rand}$)
    if refinementControl($\mathcal{T}'$, $q'_{near}$, $q_{rand}$) then
        $q_{new} \leftarrow$ extend($q'_{near}$, $q_{rand}$, $\delta$)
        if $q_{new} \neq null$ and
        transitionTest($\mathcal{T}'$, $E(q'_{near})$, $E(q_{new})$) then
            addNewNode($\mathcal{T}'$, $q_{new}$)
            addNewEdge($\mathcal{T}'$, $q'_{near}$, $q_{new}$)
            ($\mathcal{T}''$, $q''_{near}$) $\leftarrow$ findNearestTree($q_{new}$)
            if distance($q_{new}$, $q''_{near}$) $\leq \delta$ then
                $\mathcal{T} \leftarrow$ merge($\mathcal{T}'$, $q_{new}$, $\mathcal{T}''$, $q''_{near}$)
                $n \leftarrow n - 1$
```

ber of refinement nodes already present in the tree is greater than a certain ratio ρ of the total number of nodes. In practice, we consider that an expansion can yield a refinement node when the distance between q_{near} and q_{rand} is less than the extension step-size δ. Another benefit of the refinement control is to limit the number of nodes in the tree and thus to reduce the computational cost of the neighbor search. Here, we set ρ to 0.1.

Multi-T-RRT

As an extension to the T-RRT algorithm, we propose a multi-tree variant: Multi-T-RRT. Instead of building a single tree rooted at some initial conformation, the idea is to build n trees rooted at n given conformations q^k_{init}, $k = 1..n$. The pseudo-code of the Multi-T-RRT is presented in Algorithm 4. At each iteration, a tree $\mathcal{T}'$ is chosen for expansion, which can simply be done in a round-robin fashion. Then, an extension is attempted toward a randomly sampled conformation q_{rand}, starting from its nearest neighbor, q'_{near} , in $\mathcal{T}'$. If the extension succeeds, the new conformation q_{new} is added to $\mathcal{T}'$, and an edge is built between q'_{near} and q_{new}. Then, after searching for the conformation q''_{near}, which is the closest to q_{new} within all trees other than $\mathcal{T}'$, if it appears that the distance between q_{new} and q''_{near} is less than or equal to the extension step-size δ, $\mathcal{T}'$ is linked to and merged with $\mathcal{T}''$, the tree to which q''_{near} belongs. In that case, the number of trees is decreased by 1. The space exploration continues until all trees are merged into a single one or another stopping condition (number of nodes, number of expansions, running time) is met.

Besides being simple, Algorithm 4 is the most efficient way to implement the Multi-T-RRT. We have compared it to other variants, trying different strategies to expand the trees and connect them. For example, expanding the trees in a round-robin fashion toward a conformation q_{rand} sampled at each iteration is more efficient than 1) expanding all trees at each iteration toward the same conformation q_{rand}, or 2) sampling q_{rand} first and then expanding the tree that is the closest to it. Attempting to link a tree to its closest neighbor after a successful expansion is more effective than connecting it to all other trees, or to some randomly chosen trees. Attempting this connection after each successful expansion works better than doing it only when the tree's bounding box increases in size.

Results

Alanine Dipeptide

As a proof of concept, we have used the Multi-T-RRT to explore the energy landscape of the alanine dipeptide, i.e. the alanine residue acetylated in its N-terminus and methylamidated in its C-terminus: Ace-Ala-Nme. Despite its small size, it is a common test-model because of its complex energy landscape characterized by several local minima (Chodera et al. 2006). Note that, since the shape of this landscape is very sensitive to the parameters of the exploration method, we do not compare our results to those available in the literature.

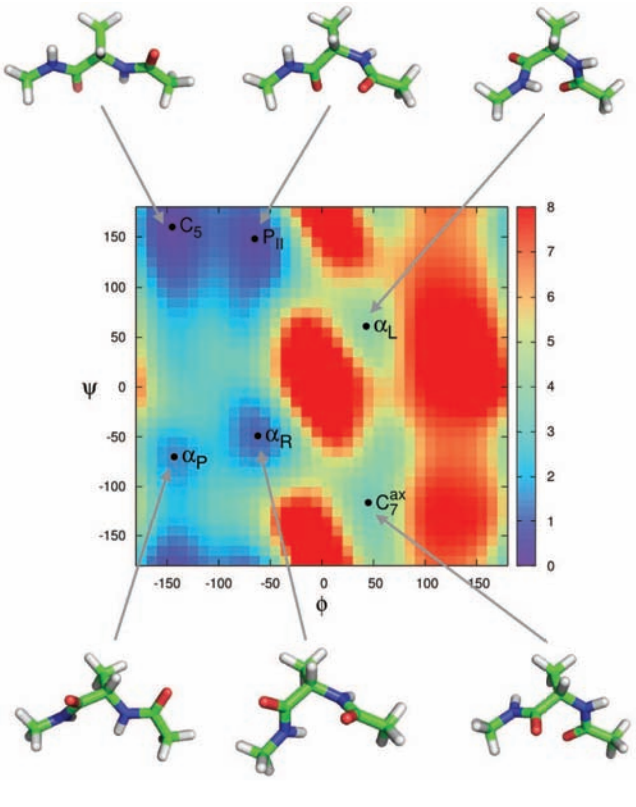

Figure 1: Ramachandran map of the alanine dipeptide in its (ϕ, ψ) space, with the locations of six local energy minima and their corresponding conformations (Chodera et al. 2006; Jaillet et al. 2011).

Table 1: Energy and (ϕ, ψ) coordinates of the local minima of the alanine dipeptide (Jaillet et al. 2011). For reference, the lowest energy value is set to 0 kcal/mol.

	C_5	P_{II}	α_R	α_P	C_7^{ax}	α_L
ϕ (°)	-145	-65	-62	-143	45	43
ψ (°)	160	148	-49	-70	-116	61
E (kcal/mol)	0	0.3	1	1.6	3.3	3.9

The conformational exploration was performed using an internal-coordinates representation of the alanine dipeptide, assuming constant bond lengths and bond angles. Therefore, the conformational parameters were the dihedral angles $\{\phi, \psi, \chi\}$ of the Ala residue, χ of the Ace capping, χ of the Nme capping, ω of the Ace-Ala peptide bond, and ω of the Ala-Nme peptide bond. As the peptide bond torsions are known to undergo only small variations, the ω angles were allowed to vary only up to 10° from the planar trans conformation.

The (ϕ, ψ) angles of the alanine dipeptide (i.e. the (ϕ, ψ) angles of the Ala residue) are very important because their flexibility allows internal hydrogen bonds to form. To visualize the results of the conformational exploration, we have used a projection of the energy landscape on these (ϕ, ψ) angles, namely the Ramachandran map. This map (see Fig. 1) was generated by varying both dihedral angles with a 10° step and energy-minimizing the conformation corresponding to each (ϕ, ψ) pair using a steepest descent method while blocking the (ϕ, ψ) angles (Jaillet et al. 2011).

The local energy minima used as input for the conformational exploration were six stable states of the alanine dipeptide (see Fig. 1), namely the C_5, P_{II}, α_R, α_P, C_7^{ax}, α_L states (Chodera et al. 2006). The conformations corresponding to these minima were produced by an iterative simulated annealing protocol (Jaillet et al. 2011). Their energy and (ϕ, ψ) coordinates are presented in Table 1.

Force Field

To compute conformational energy values, we have used the AMBER parm96 force-field with an implicit representation of the solvent using the Generalized Born approximation. Note that, for the sake of computational efficiency, we have implemented our own version of this force field as part of our application. This avoids having to make system calls to the AMBER tools.

Multi-T-RRT Parameters

The conformations used as input for the Multi-T-RRT were the six aforementioned local energy minima. The conformational distance required by the Multi-T-RRT was defined as the Euclidean distance in the (ϕ, ψ) space. The extension step-size δ was set to 0.1, so that the maximal angular variation between two conformations was about 6°. The (relative) energy threshold E_{max} in the transition test was set to 8 kcal/mol (see the energy scale on the right-hand side of the Ramachandran map). The Boltzmann constant being $1.987 \cdot 10^{-3}$ kcal/mol/K, by setting the initial temperature to 70 K, we imposed the probability of accepting an energy increment of 0.1 kcal/mol to be around 50% at the beginning of the exploration.

Transition Paths

The Multi-T-RRT was used to compute transition paths between the local energy minima of the alanine dipeptide. To get an idea of the likelihood of the produced transition paths,

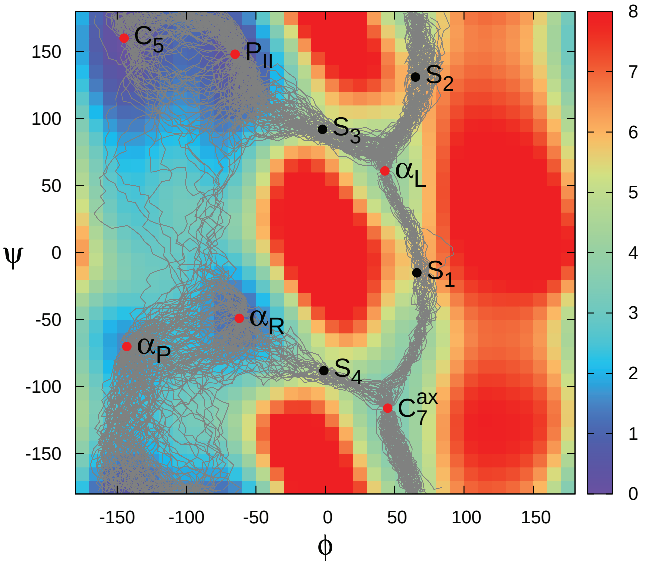

Figure 2: Projection on the (ϕ, ψ) space of the alanine dipeptide of the transition paths produced by 100 runs of the Multi-T-RRT, with $T_{rate} = 0.1$.

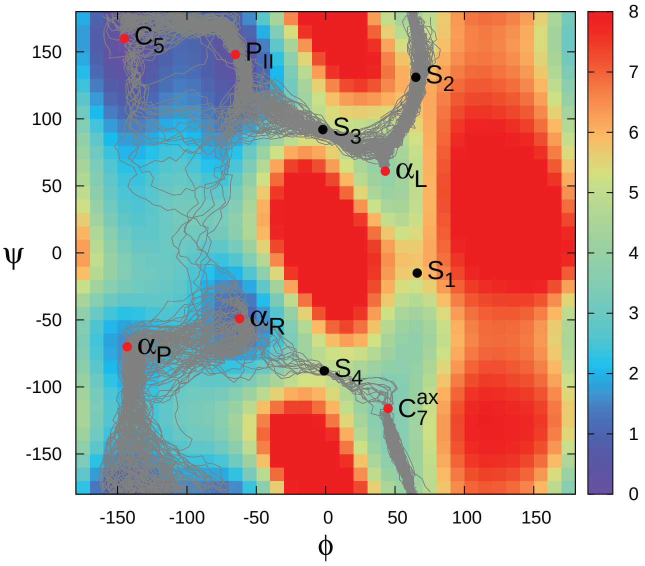

Figure 3: Projection on the (ϕ, ψ) space of the alanine dipeptide of the transition paths produced by 100 runs of the Multi-T-RRT, with $T_{rate} = 0.01$.

the Multi-T-RRT was run 100 times. After each run, i.e. when the six trees rooted at the different minima were all merged into a single tree, a path was extracted from that tree for each pair of minima and projected on the (ϕ, ψ) space. Fig. 2 and Fig. 3 were obtained by aggregating the results of these 100 runs of the Multi-T-RRT with $T_{rate} = 0.1$ and $T_{rate} = 0.01$ respectively. The first outcome of these tests is that the multi-T-RRT is extremely fast: Fig. 2 was produced in about 1 min (i.e. less than 1 s for each run), and Fig. 3 was produced in about 8 min (i.e. about 5 s for each run).

As already mentioned, when $T_{rate} = 0.1$, the Multi-T-RRT covers the conformational space more quickly than when $T_{rate} = 0.01$. But, in the latter case, it produces lower-energy transition paths. In fact, the regions containing the transition paths in Fig. 3 are narrower and fit better within the lower-energy areas of the landscape, in comparison to Fig. 2. Moreover, when $T_{rate} = 0.1$, some transition paths between α_L and C_7^{ax} go through the saddle point S_1, whereas, when $T_{rate} = 0.01$, all transition paths between α_L and C_7^{ax} go through the saddle point S_2 whose energy is lower (see Table 2). As a general remark, we observe that none of the transition paths goes through the energetic barrier corresponding to $\phi \in [100°, 150°]$. Finally, only few transition paths go through the medium-energy area corresponding to $\phi \in [-150°, -50°]$ and $\psi \in [0°, 50°]$.

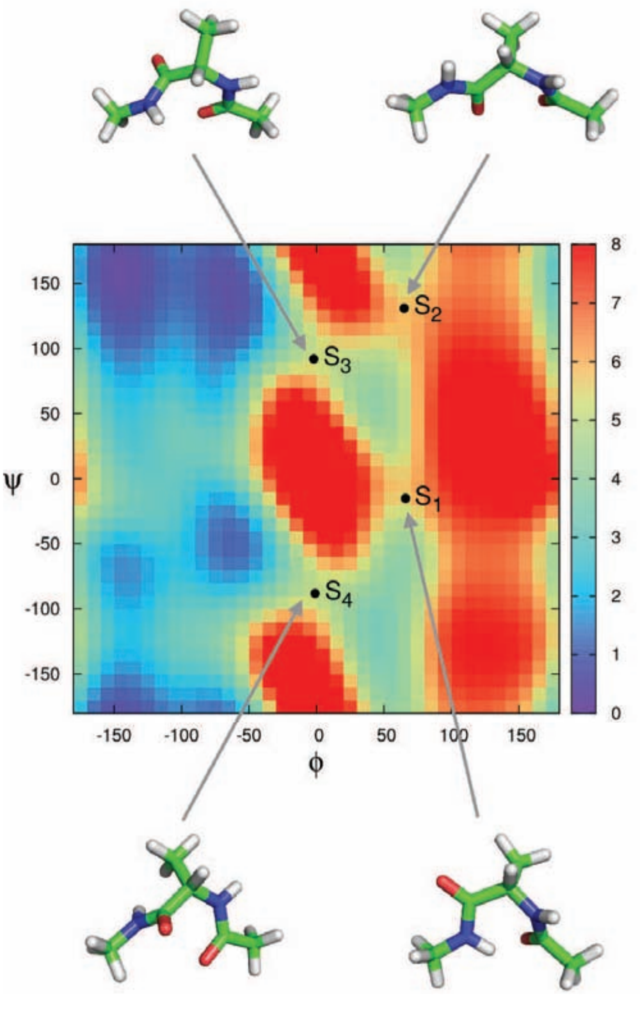

Figure 4: Locations and conformations of the four main saddle-points of the alanine dipeptide.

Table 2: Relative energy and (ϕ, ψ) coordinates of the saddle-points of the alanine dipeptide.

	S_1	S_2	S_3	S_4
ϕ (°)	66	65	-2	-1
ψ (°)	-15	131	92	-88
E (kcal/mol)	5.87	5.85	5.21	5.03

Saddle Points

We have also used the Multi-T-RRT to find the main transition states (i.e. the saddle points) of the alanine dipeptide. For that, we have run the Multi-T-RRT 1000 times (500 times with $T_{rate} = 0.1$ and 500 times with $T_{rate} = 0.01$). After each run, we have extracted from the produced tree the transition paths between α_L and C_7^{ax}, between α_L and P_{II}, and between α_R and C_7^{ax}; we have also computed the maximal energy observed along each of these paths. Then, for each class of transition path, we have extracted the conformation having the lowest energy maximum, across the 1000 runs. Following this procedure, we have isolated the four saddle points shown in Fig. 4. Their energy and (ϕ, ψ) coordinates are presented in Table 2.

Conclusion

We have addressed the problem of computing transition paths between stables states of a molecule by exploring its energy landscape. We have based our work on the use of the T-RRT algorithm, which originates from the robotics field, but which has been already used to sample the conformational space of a molecule. In this paper, we have proposed a multi-tree extension of T-RRT: the Multi-T-RRT. Instead of exploring the conformational space by growing a single tree, the Multi-T-RRT constructs several trees rooted at different interesting conformations. We have evaluated this algorithm on the alanine dipeptide, by computing transition paths between local minima of its energy landscape, as well as their associated saddle-points. The main benefit of the Multi-T-RRT is to quickly provide some interesting information about the regions through which these transition paths might go. Indeed, a single run of the Multi-T-RRT can terminate within seconds, and it takes only a few minutes to run it several times and aggregate the results.

This work is only a preliminary step in the direction of what we plan to achieve. As future work, we aim to develop a version of the Multi-T-RRT that could produce in a single run some transition paths that could have a relevant interpretation, which is not the case now (because we have to aggregate the results produced by several runs before interpreting them). For that, instead of building a tree obtained by growing and merging several trees over the conformational space, we would have to build a graph potentially containing cycles. This would allow the Multi-T-RRT to produce more interesting transition paths through the energy landscape. Then, we could evaluate the algorithm on other, more complex, systems, and compare our results to those obtained with other methods.

Acknowledgments

This work has been partially supported by the French National Research Agency (ANR) under project ProtiCAD (project number ANR-12-MONU-0015-01).

References

Al-Bluwi, I.; Siméon, T.; and Cortés, J. 2012. Motion planning algorithms for molecular simulations: A survey. *Computer Science Review* 6(4):125–143.

Chodera, J. D.; Swope, W. C.; Pitera, J. W.; and Dill, K. A. 2006. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling and Simulation* 5(4):1214–1226.

Frenkel, D., and Smit, B. 2001. *Understanding Molecular Simulations: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press, 2nd edition.

Gipson, B.; Hsu, D.; Kavraki, L. E.; and Latombe, J.-C. 2012. Computational models of protein kinematics and dynamics: Beyond simulation. *Annual Review of Analytical Chemistry* 5:273–291.

Jaillet, L.; Corcho, F. J.; Pérez, J.-J.; and Cortés, J. 2011. Randomized tree construction algorithm to explore energy landscapes. *Journal of Computational Chemistry* 32(16):3464–3474.

LaValle, S. M., and Kuffner, J. J. 2001. Rapidly-exploring random trees: Progress and prospects. In *Algorithmic and Computational Robotics: New Directions*. A K Peters.

Olson, B. S., and Shehu, A. 2012. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Science* 10(Suppl 1):S5.

Wales, D. J., and Doye, J. P. K. 1997. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A* 101(28):5111–5116.

Wales, D. 2003. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press.

Wilson, S. R.; Cui, W.; Moskowitz, J. W.; and Schmidt, K. E. 1991. Applications of simulated annealing to the conformational analysis of flexible molecules. *Journal of Computational Chemistry* 12(3):342–349.