A Randomized Tree Construction Algorithm to Explore Energy Landscapes

L. Jaillet[1], F.J. Corcho[2], J.J. Pérez[2], J. Cortés[3,4]

[1] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, C/ Llorens i Artigas 4-6, Barcelona 08028, Spain. Email: ljaillet@iri.upc.edu

[2] Dep. d'Enginyeria Química, UPC. ETS d'Enginyeria Industrial, Av. Diagonal, 647, 08028 Barcelona, Spain. Email: {francesc.corcho, juan.jesus.perez}@upc.edu

[3] CNRS; LAAS; 7 avenue du colonel Roche, F-31077 Toulouse Cedex 4, France. Email: juan.cortes@laas.fr

[4] Université de Toulouse; UPS, INSA, INP, ISAE; UT1, UTM, LAAS; F-31077 Toulouse Cedex 4, France.

Abstract

We report in the present work a new method for exploring conformational energy landscapes. The method, called T-RRT, combines ideas from statistical physics and robot path planning algorithms. A search tree is constructed on the conformational space starting from a given state. The tree expansion is driven by a double strategy: on the one hand, it is naturally biased towards yet unexplored regions of the space; on the other, a Monte Carlo-like transition test guides the expansion toward energetically favorable regions. The balance between these two strategies is automatically achieved thanks to a self-tuning mechanism. The method is able to efficiently find both, energy minima and transition paths between them. As a proof of concept, the method is applied to two academic benchmarks and to the alanine dipeptide.

Keywords : energy landscape exploration, robot path planning algorithms, Monte Carlo methods, conformational transition paths, peptides.

1 Introduction

There is a strong body of evidence that the physicochemical properties of molecules are related to their atomic structure. Since structural-dependent molecular properties can not always be determined by experimental methods, computational methods represent a reliable alternative to obtain properties of matter at the molecular level. The accurate prediction of molecular properties with computational methods requires an adequate sampling of the states of interest. This is a challenging task since the number of states grows exponentially with the size of the system, hampering the performance of search methods, and consequently, the accuracy of the predictions. The problem of performing a significant search of the molecular energy landscape has attracted the interest of the scientific community for decades [1, 2, 3]. Basically, there are two approaches to tackle the problem: i) to sample the conformational space, producing Boltzmann-weighted ensembles; ii) to characterize stationary points and transition paths on the conformational energy surface. In the former category, methods like the Metropolis Monte Carlo (MC) or Molecular Dynamics (MD) [4, 5] are used to compute the thermodynamic properties of a system and, in the case of MD, also its dynamical properties. The drawback of these methods is the long sampling/simulation times required to surmount energy barriers that separate relevant conformations and perform an efficient sampling. In order to overcome this difficulty, methods such as the replica exchange [6, 7], umbrella sampling [8, 9], the activation-relaxation technique [10], or metadynamics [11] have been designed to bias the sampling process, enhancing it on specific degrees of freedom or infrequent events [12]. The latter category of methods includes algorithms that explore the topography of the conformational energy surface aiming to find energy minima corresponding to stable states and probable transition paths between such states. Methods to find energy minima [13] usually combine conformational sampling and energy minimization. The most widely used methods are based on genetic algorithms [14], the simulated annealing procedure [15], the basin hopping strategy [16], and taboo search [17]. Very diverse methods have been proposed to find transition paths between two given stable states. In general, the idea is to start from a trivial path and to deform it locally and iteratively in order to improve its energy profile. Examples of such methods are the nudged elastic band [18, 19], the zero-temperature string method [20], and its extension for rough energy landscapes: the finite-temperature string method [21]. A recent approach called forward flux sampling [22] does not require an initial path, but needs to define series of interfaces between the initial state and the target state. An alternative method is transition path sampling [23], which develops MC specific procedures to produce a set of reactive trajectories describing the dynamical pathways that bridge stable states. Methods for computing conformational transitions have also been developed based on biased or targeted MD [24, 25] and on normal mode analysis [26]. Finally mention a category of methods, including the lid algorithm [27] (for discrete spaces) and the threshold algorithm [28] (for continuous spaces), originally devoted to explore exhaustively the regions around a given set of energy minima, and which, by extension, are able to characterize conformational transitions between minima without prior information about their connectivity.

All the aforementioned methods present specific advantages and drawbacks, which make them more or less suitable to particular applications. Nevertheless, there is still room for the development of efficient and general methods to explore energy landscapes. Recent works show that algorithms originating from robotics can be the basis for the development of efficient conformational sampling and exploration methods in computational biochemistry. For instance, methods based on robotics algorithms have been proposed to analyze protein loop mobility [29, 30], to compute large-amplitude conformational transitions in proteins [31, 32], to investigate protein and RNA folding pathways [33, 34], or to simulate ligand diffusion inside proteins considering flexible molecular models [35, 36]. The present work proposes a conformational exploration method, called Transition-RRT (T-RRT) [37], which is inspired by robotic path planning algorithms and by methods in statistical physics. T-RRT can be seen as a non-canonical sampling method to identify interesting points on the energy landscape (i.e. minima and saddle-points) and/or as a method to compute energetically favorable conformational transition paths. Similarly to MC methods, T-RRT applies small moves to the system and uses a probability transition test based on the Metropolis criterion. However, instead of generating a single path on the conformational

space, it constructs a tree with a specific expansion mechanism that provides better coverage properties. Such a data structure enables the simultaneous exploration of different regions of the space. Moreover, in contrast to MC methods, it avoids to waste time getting back to regions of the space already explored. Finally, T-RRT is a reactive search method [38] that uses a self-tuning mechanism to improve its overall efficiency. Starting from a given conformation, the tree branches grow first on the more favorable regions (the valleys of the landscape), and tend to cover the whole search-space while the number of iterations increases. Such exploration enables to find the local minima and the saddle-points of the landscape. Furthermore, paths extracted from the tree can be directly exploited as a good approximation of transition paths between stable conformations.

2 Methods

This section describes the Transition-RRT algorithm (T-RRT), whose pseudo-code is sketched in Algorithm 1. T-RRT extends the Rapidly-exploring Random Tree (RRT) algorithm [39] by incorporating a stochastic state-transition test, similarly to MC methods. RRT is a randomized space-filling method that was initially developed for path planning in robotics. Its most interesting feature is the implicit bias of the tree expansion toward yet unexplored regions of the space.

The proposed variant, T-RRT, also holds this interesting property. In addition, it integrates a transition test that filters some of the states generated when they do not correspond to energetically acceptable moves. Thus, the expansion is biased toward both unexplored and low energy regions. The appropriate balance between these two types of bias relies on a reactive scheme as described below. Also, a filtering procedure rejects new states if they are too close to states already stored within the tree, which improves the space-covering property of the method. Overall, T-RRT is an effective and general exploration method that can be used to find stable states, or to compute probable transition paths between given pairs of states.

2.1 RRT Principle: Bias Toward Unexplored Regions

The core of the T-RRT algorithm (Algorithm 1) is inherited from the basic RRT [39]. RRT is an efficient path planning method able to tackle complex problems in high-dimensional spaces. It has been successively used in several disciplines such as robotics, computer animation, and computational biochemistry [35, 32]. The idea is to iteratively construct a tree data structure made of nodes and edges that correspond to states and small-amplitude motions between neighbor states, respectively. At each iteration, a state (i.e. a molecular conformation in the present context) is randomly sampled (`SampleConf` function). The nearest state already contained in the search tree is then searched (`NearestNeighbor` function). Finally, a new node is created by extending the nearest neighbor toward the random sample (`Extend` function). Employing the simplest expansion strategy (called RRT-Extend in related literature [39]), the extension step-size δ remains constant for all the iterations. The main interest of such a construction procedure (illustrated in Figure 1) is that the tree expansion is implicitly biased toward yet unexplored regions. This behavior comes from the probability for a node to be extended, which is proportional to the volume of its Voronoi cell (i.e. the set of points closer to this node than to any other node). Note that this property does not require the explicit construction of the Voronoi cells, which would be computationally expensive.

Classically, RRT has been used to search paths in a continuous state-space composed of feasible-state and unfeasible-state subsets. In this context, RRT has been proved to be much more efficient than a simple random walk method, since it avoids wandering around in already explored regions [40]. By inference, T-RRT is expected to be more effective than standard MC methods to explore molecular energy landscapes.

2.2 Transition Test: Hindering Steep Climbing

T-RRT extends RRT by integrating a transition test to hinder the tree expansion toward energetically unfavorable regions of the space (`TransitionTest` function). Similarly to MC methods, the acceptance rule of a local move is defined by comparing the energy E_j of the new state with the energy E_i of the previous state (i.e. the parent node in the tree). This test is based on the Metropolis criterion, with a transition probability p_{ij} defined as follows:

$$p_{ij} = \begin{cases} exp(-\frac{\Delta E_{ij}}{kT}), & \text{if } \Delta E_{ij} > 0 \\ 1, & \text{otherwise} \end{cases}$$

where $\Delta E_{ij} = E_j - E_i$ is the energy variation between the two states, k is the Boltzmann constant, and T is the temperature. Note however that T-RRT is a non-canonical sampling method, which is not expected to produce a Boltzmann weighted set of conformations, but to efficiently find energy minima and probable conformational transition paths. Consequently, T can be considered as a parameter of the algorithm, and does not necessarily carry any physical meaning.

Within search methods involving the Metropolis criterion, the temperature is usually kept constant (e.g. MC simulation) or is subject to predefined variations (e.g. heating and cooling phases in simulated annealing). In the case of T-RRT, the `TransitionTest` function incorporates a reactive scheme to dynamically tune this parameter. It allows controlling the level of difficulty of the transition test, according to the information acquired during the exploration.

2.3 Automatic Temperature Tuning

During the construction of the search tree, the number of attempts necessary to add a new node is a good indicator to measure the evolution of the exploration process. A large number of consecutive transition failures means that the exploration is stuck because the tree cannot be further expanded toward favorable regions. Within T-RRT, this information is used to regulate the temperature that determines the difficulty of the transition test.

At the initialization, T is set to a low value in order to only permit the tree expansion on very easy positive slopes (in addition to flat and negative ones). Then, during the exploration, the number of consecutive times the Metropolis criterion discards a state is recorded and used for temperature tuning. When the T-RRT search reaches a maximum number of consecutive rejections $Fail_{max}$, the temperature increases by a factor λ, which increases the probability to succeed the transition test in subsequent iterations. Contrarily, each time an uphill transition test succeeds, the temperature decreases by the same factor λ, therefore increasing the severity of the transition test. Thus, the temperature is automatically regulated along the exploration in function of the energy landscape profile. This temperature regulation strategy is a way to balance the search between unexplored regions and low energy regions.

2.4 Exploration Guarantee

The adaptive temperature tuning introduced above may however lead to bottleneck situations. The temperature T may be reduced by the insertion of new states very similar to the ones already contained in the tree, whereas the expansion toward new regions of the space would require an increment of T. The insertion of such states only contributes to the refinement of the exploration in regions already reached by the tree. This situation is illustrated on a 2D fictive energy landscape in Figure 2-a.

To overcome this drawback, the selected state q_{near} is not extended if the distance to the random state q_{rand} is smaller than the extension step-size δ (`ExploGuarantee` function). Such a simple filtering avoids an excessive refinement of low-energy regions, therefore facilitating the tree expansion toward new regions of the space. Furthermore, it limits the size of the tree (in number of nodes), which reduces the computational cost of operations such as neighbor search. The improvement provided by this filtering process is illustrated in Figure 2-b.

2.5 T-RRT Operating Modes

2.5.1 Main Minima Search Method

The proposed method can be used to find the main minima of a conformational landscape. Hereafter, this type of operating mode is called T-RRT$_{min}$. Starting from a given conformation, the method builds a tree that explores the landscape until a stop condition is reached. This condition can be defined by an amount of computing time, a maximum number of created conformations, or from an estimation of the space coverage. Once the search is stopped, a minima extraction method can be applied to the conformations contained in the tree. In the current implementation, we apply a method based on the the root mean square deviation (RMSD) between conformations: the main minima are the conformations whose energy is lower than the energy of all their neighbors for a given RMSD threshold.

2.5.2 Transition Path Search Method

The previous application, for which other effective methods are available (e.g. [16, 17]), is nevertheless not the most suitable use of T-RRT. Rather, T-RRT is particularly well suited to find low-energy paths between a given pair of stable conformations, and to identify the transitions states associated with. This operating mode is called T-RRT$_{trans}$. For such a search, the tree is rooted at one of the stable conformations, and the algorithm is iterated until one of the tree leaves reaches the target conformation (i.e. the distance between both conformations is less than the extension step-size δ). The transition path is then extracted from the tree structure, by following the branches from the leaf to the root. The quality of the computed path relies on two points. First, the Voronoi bias avoids backtrack motions, contrarily to basic MC techniques that propagate a single state. Second, the temperature is regulated all over the tree construction process, so that heating phases only occur when necessary for passing through higher energy barriers in order to reach other conformational regions. Consequently, paths computed by T-RRT tend to minimize the total amount of positive energy variation (empirical proofs have been provided by Jaillet *et al.* [37]). Therefore, such paths are good candidates to represent transitions between pairs of stable conformations.

The approach could be extended to find connections between a given set of n minima. This will require the implementation of a multi-tree variant of T-RRT, sharing ideas with methods that generalize the basic RRT algorithm to multiple trees [41]. A tree could be constructed to explore the space around each of the n minima. Connections between leaves of nearby trees will permit to identify possible conformational transitions between minima. The result will be a graph of transition paths between the set of n minima. Such a possible extension of T-RRT presents similarities with the threshold algorithm [28]. Note however that previous work [37] shows that T-RRT outperforms a method that introduces a threshold-based strategy within the RRT algorithm [42] for robot path planning on rough terrains. Nevertheless, the implementation of such a multi-tree variant of T-RRT, as well as its comparison with related methods in computational chemistry, remain for future work.

3 Results

As a proof of concept, this section first presents results on two academic benchmarks, for which the energy landscape is represented by a two-parameter analytic function. Then, the method is applied to study the energy landscape of the alanine dipeptide using an implicit description of the solvent.

For each problem, T-RRT is first used to find the main energy minima (T-RRT$_{min}$ search), and then to find the transition paths between these states (T-RRT$_{trans}$ search). The algorithm parameters are the following. The Boltzmann constant k being $3.297 \cdot 10^{-27}$ kcal/K, the initial temperature is set to $T = 70$ K. This value imposes that, at the initialization of the algorithm, the probability of accepting an energy increment of 0.1 kcal/mol is around 50%. The maximum

number of consecutive expansion failures before a temperature increase is set to $Fail_{max} = 10$ and $Fail_{max} = 100$ for T-RRT$_{min}$ and T-RRT$_{trans}$, respectively. With these settings, T-RRT$_{min}$ covers the space more rapidly than T-RRT$_{trans}$, while T-RRT$_{trans}$ finds the saddle-points more accurately. The temperature variation factor is $\lambda = 0.1$ in all the cases.

3.1 2D Academic Benchmarks

The apparently simple landscapes[1] described below represent tricky test systems for benchmarking methods that search for conformational transition pathways.

The *Zorro* potential, represented in Figure 3, involves two low-energy regions with respective minima A and B. The pathway connecting these minima needs to circumvent two energy barriers and passes through a saddle-point located in the middle part of the landscape. The analytic expression of this energy landscape is:

$$\begin{aligned} E(x,y) \;=\; & 0.2((x/8)^4 + (y/8)^4) \\ & - 3e^{-0.2(0.05(x+5)^2+(y+5)^2)} - 3e^{-0.2(0.05(x-5)^2+(y-5)^2)} \\ & + 5e^{-0.2(x+3(y-3))^2}/(1+e^{-x-3}) \\ & + 5e^{-0.2(x+3(y+3))^2}/(1+e^{x-3}) + 3e^{-0.01(x^2+y^2)} \\ & + 0.06 * (\sin(5x+\sqrt{2}y) + \cos(\sqrt{5}x+\sqrt{3}y) \\ & + \sin(3*y-\sqrt{2}x) + \cos(3*x-\sqrt{5}y)). \end{aligned}$$

The variation of parameters x and y is limited to the interval $[-15, 15]$. Within these bounds, the energy varies from 0.5 up to 9.7 (in arbitrary units).

The *Alien* potential is represented in Figure 4. Like the previous benchmark, it involves two low-energy basins with respective minima A and B. These regions are connected through two main pathways, which we refer to as lower and upper (l and u in Figure 4). The energy value of the transition states for these two pathways is very similar. However, the upper pathway is much larger than the lower one. The analytic expression of this energy landscape is given by:

$$\begin{aligned} E(x,y) \;=\; & 3 + \frac{3}{e^{5(\frac{x^2}{4}+\frac{(2+\frac{y}{2})^2}{10})}} - \frac{3}{e^{5\left((-2+\frac{x}{2})^2+(2+\frac{y}{2})^2\right)}} \\ & - \frac{3}{e^{5\left((2+\frac{x}{2})^2+(2+\frac{y}{2})^2\right)}} + \frac{(\frac{x^2}{4}+\frac{y^2}{8})^4}{10000} + \frac{1+erf(1+\frac{y}{2})}{2} \\ & + \frac{3}{50}\left(\cos(\sqrt{5}x+\sqrt{3}y) + \cos(3x-\sqrt{5}y))\right) \\ & + \frac{3}{50}\left(\sin(5x+\sqrt{2}y) - \sin(\sqrt{2}x-3y)\right). \end{aligned}$$

Like in the previous example, the variation of parameters x and y is limited to the interval $[-15, 15]$. Within these bounds, the energy varies from 0.1 up to more than 5000 (in arbitrary units).

The low-dimensionality of these benchmarks enables comparison of results with those obtained by exhaustive search. A 128×128 grid discretizing the search-space was used to perform such a search. In order to analyze the variability of T-RRT results (due to the randomized exploration), the algorithm was run several times on each problem. Results presented below show the average and the standard deviation over 100 runs.

3.1.1 T-RRT$_{min}$ Search

Table 1 shows results of the T-RRT$_{min}$ search for the *Zorro* and the *Alien* benchmarks. In both cases, the minima found are very similar (in position as well as in energy) to those extracted from an exhaustive grid search method. Moreover, the low values of the standard deviation confirm the

Zorro

	T-RRT$_{min}$		Grid	
	A	B	A	B
x	7.6 ± 0.2	-5.6 ± 0.1	7.7	-5.6
y	-5.1 ± 0.2	5.1 ± 0.1	-5.1	5.1
E	0.59 ± 0.03	0.60 ± 0.02	0.55	0.57

Alien

	T-RRT$_{min}$		Grid	
	A	B	A	B
x	-4.0 ± 0.2	4.1 ± 0.2	-3.9	4.1
y	-4.0 ± 0.2	-4.1 ± 0.2	-4.1	-3.9
E	0.30 ± 0.18	0.51 ± 0.18	0.14	0.34

Table 1: Energy minima for the academic benchmarks.

Zorro

	T-RRT	Grid
	A $\rightarrow$ B	A $\rightarrow$ B
x	-0.9 ± 0.6	0.5
y	0.7 ± 0.6	-0.6
E	4.96 ± 0.02	4.95

Alien

	T-RRT		Grid	
	A $\xrightarrow{l}$ B	A $\xrightarrow{u}$ B	A $\xrightarrow{l}$ B	A $\xrightarrow{u}$ B
x	0.2 ± 0.1	0.8 ± 2.3	0.1	-0.9
y	-8.6 ± 0.2	3.3 ± 2.4	-8.2	0.5
E	3.98 ± 0.03	4.13 ± 0.08	3.96	4.07

Table 2: Transition states for the academic benchmarks.

reliability of T-RRT despite the random nature of the search process. For these experiments, the T-RRT iteration was stopped after the insertion of 1000 nodes. The extension step-size (i.e. the Euclidean distance between two connected states in the tree) was set to $\delta = 0.5$.

3.1.2 T-RRT$_{trans}$ Search

The localization of the minima being worked out, T-RRT was used to find transition paths between them. Table 2 shows the position and the energy of the associated transition states found for the two benchmarks. In both cases, the transition states computed with T-RRT are very close to the ones found by exhaustive search. Moreover, in the case of the *Alien*, both the lower and the upper pathways were found with T-RRT. Over the 100 computed paths, 66 passed through the upper region whereas only 34 passed through the lower one. This result shows that, for energy barriers of similar height, T-RRT exhibits a higher probability to pass through wider transition regions than through narrow passages. In other words, T-RRT solutions do not only depend on the potential energy of the explored states, but also on the number of possible equivalent paths to reach the other minimum. Consequently, results indicate that the method tends to search for the lowest free energy route.

	P_{II}	α_R	α_L	C_7^{ax}	α_P	C_5
ϕ	-67	-63	47	50	-148	-146
ψ	144	-44	51	-138	-70	162
E	0.3	1.1	4.4	4.2	1.7	0.0

Table 3: Energy minima of alanine dipeptide obtained by T-RRT.

3.2 Alanine Dipeptide in Implicit Solvent

The alanine dipeptide refers to the alanine residue acetylated in its N-terminus and methylamidated in its C-terminus (see Figure 5). It is a relatively small biomolecule with a complex energy landscape characterized by several local minima and intermediates connected by multiple pathways, being a frequent test-model molecule for theoretical studies [43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53]. Despite its small size, alanine dipeptide shares some structural features with larger peptides and proteins. In particular, due to the flexibility of the ϕ and ψ angles, the molecule is able to form internal hydrogen bonds. However, it should be noted that the overall shape of the conformational energy landscape of alanine dipeptide (i.e. the number of minima, the energy of transition states, ...) is very sensitive to simulation conditions [49, 53]. Such a sensitivity hampers a direct comparison of results to those available in the literature. In this work, we have used the AMBER parm96 force-field [54] together with an implicit representation of the solvent using the Generalized Born approximation for convenience. The values of the internal and external dielectric constants were set to 1.0 and 78.5, respectively.

For facilitating the analysis of results obtained with T-RRT, an energy map on the $\{\phi, \psi\}$ coordinates of the peptide (i.e. the Ramachandran map) was generated using a systematic procedure. The two dihedral angles were varied with constant 10° step-size. For each $\{\phi, \psi\}$ value, the conformation was energy-minimized using a steepest descent method. In order to fix the $\{\phi, \psi\}$ angles during the minimization, we used an additional $\{\phi, \psi\}$-harmonic potential whose minimum was equal to the desired values of the two angles. The optimization was stopped when the RMSD for consecutive iterations reached $1 \cdot 10^{-3}$Å. The computed energy map appears in background in Figures 6 to 10.

The conformational exploration with T-RRT was performed on an internal-coordinate representation of alanine dipeptide with constant bond lengths and bond angles. Thus, the conformational parameters are the seven bond torsions associated with the dihedral angles ϕ, ψ, $\omega_{1,2}$, and $\chi_{1,2,3}$ represented in Figure 5. Note that, since the peptide bond torsions $\omega_{1,2}$ are known to undergo small variations, they were allowed to vary only $\pm 10°$ from the planar trans conformation.

3.2.1 T-RRT$_{min}$ Search

The energy landscape exploration with T-RRT yielded six minima that correspond to the P_{II}, α_R, α_L, C_7^{ax}, α_P and C_5 stable states of the alanine dipeptide [48]. Their position and energy are presented in Table 3 (for reference, the energy of the minimum-energy conformation is set to zero). Figure 6 shows these minima projected on the $\{\phi, \psi\}$ energy map. It appears that T-RRT solutions fit very well the minimum energy regions of the map. This result shows the capacity of the method to find multiple minima in multidimensional landscapes. For obtaining these results, T-RRT$_{min}$ was iterated until the insertion of 8000 nodes in the tree. The exploration step-size δ was set such that the maximal angular variation was of 5°. Once the tree was constructed, the six minima were identified by applying the minima extraction method based on RSMD, described in Section 2.5. Finally, these minima were locally optimized by a steepest descent method with the same stop criterion than that used for the construction of the $\{\phi, \psi\}$ energy map.

For comparison, the search of minima was also performed using an iterative simulated annealing (SA) protocol [55]. The initial minimized structure was quickly heated up to 900 K at a rate of 100 K/ps, in order to force the molecule to jump to a different region of the conformational space. Subsequently, the 900 K structure was slowly cooled to 200 K at a rate of 7 K/ps and then minimized. The minimization was carried out with a steepest descent algorithm and with the

	P_{II}	α_R	α_L	C_7^{ax}	α_P	C_5
ϕ	-65	-62	43	45	-143	-145
ψ	148	-49	61	-116	-70	160
E	0.3	1.0	3.9	3.3	1.6	0.0

Table 4: Energy minima of alanine dipeptide obtained by a simulated annealing protocol.

	S_1	S_2	S_3	S_4	S_5	S_6
ϕ	0	3	72	74	-111	-142
ψ	95	-90	137	-8	10	-118
E	7.3	7.7	7.3	7.7	3.4	2.6

Table 5: Main transition states of alanine dipeptide found by T-RRT.

convergence criterion set to 0.001 kcal·mol^{-1}·Å^{-1}. The structure obtained at $200K$ was stored and used as the starting conformation for a new cycle of SA. The SA procedure was run for 2000 cycles, which yielded a list of 2000 structures that was ordered by energy. These conformations were then checked for uniqueness, and those for which none of the backbone dihedral angles were different from at least 60° with respect to lower-energy conformations were excluded from the list. This process yielded 6 energy minima, whose values and positions are presented in Table 4. These results are very similar to those obtained by T-RRT$_{min}$. The differences are slightly more significant for the two minima with higher energy: C_7^{ax} and α_L. One explanation for these minor differences would come from the use of different coordinates to perform the search with SA (Cartesian coordinates) and T-RRT (internal coordinates with fixed bond lengths and bond angles). Finally mention that the computing time required by the SA protocol for finding the 6 minima was of 660 minutes, whereas the T-RRT search took only 15 minutes using the same force-field and solvent model.

3.2.2 T-RRT$_{trans}$ Search

T-RRT was used to compute transition paths between several pairs of minima. For facilitating the analysis of results, the positions and energy of the main transition states (saddle points), extracted from the total set of computed paths, are presented in Table 5, and they are located on the map in Figure 6. The algorithm was run 100 times for each transition in order to perform a statistical analysis on the probabilities associated with different classes of transition paths. However, only the first 50 computed paths are shown in the figures for clarity of presentation.

Figure 7 and Figure 8 show transition paths for the direct and reverse transition between pairs of minima $\{\alpha_L, C_7^{ax}\}$ and $\{\alpha_R, P_{II}\}$, respectively. These transitions, that are known to be fast transitions [52], do not require to cross the lines $\phi = 0°$ or $\phi = 120°$. They mainly involve the variation of bond torsion ψ, and the paths lead from one minima to the other via a single saddle-point in all the cases. In the case of $\alpha_L \leftrightarrow C_7^{ax}$ transitions, class-I paths go across S_3, while class-II paths go across S_4. The number of paths of each class found after 100 runs of T-RRT$_{trans}$ are given in Table 6. The higher probability to find class-I paths, for both the direct $\alpha_L \rightarrow C_7^{ax}$ transition and the reverse $C_7^{ax} \rightarrow \alpha_L$ transition, can be explained by a lower energy barrier to cross S_3 than to cross S_4. The variability of the paths within each class is very low, due to the narrowness of the corridors that connect the two minima. Two classes of transition paths are also found between α_R and P_{II}. However, contrarily to $\alpha_L \leftrightarrow C_7^{ax}$ transitions, the distribution of the paths depends on the direction (see Table 6). Class-II paths are more frequent than class-I paths for the $\alpha_R \rightarrow P_{II}$ transition, whereas the probabilities are inverted for the reverse $P_{II} \rightarrow \alpha_R$ transition. Such a different behavior can be explained through an analysis of the local topography of the landscape around these minima. Paths starting from α_R have a higher tendency to reach α_P (α_R and α_P are separated by a relatively low energy barrier that is not indicated in the figures) than to reach directly the P_{II}/C_5 region going across S_5. From α_P, the P_{II}/C_5 region is reached passing through S_6, which requires lower energy than crossing S_5. On

	I	II
$\alpha_L \rightarrow C_7^{ax}$	62%	38%
$C_7^{ax} \rightarrow \alpha_L$	60%	40%
$\alpha_R \rightarrow P_{II}$	34%	66%
$P_{II} \rightarrow \alpha_R$	64%	36%

Table 6: Distribution of T-RRT solutions into path classes for conformational transitions of alanine dipeptide that do not require to cross $\phi = 0°$ or $\phi = 120°$.

	I	II	III	IV	V	VI
$\alpha_R \rightarrow C_7^{ax}$	21%	54%	2%	8%	2%	13%
$C_7^{ax} \rightarrow \alpha_R$	9%	17%	31%	27%	11%	5%
$C_5 \rightarrow C_7^{ax}$	33%	16%	2%	34%	1%	14%
$C_7^{ax} \rightarrow C_5$	53%	9%	32%	1%	4%	1%

Table 7: Distribution of T-RRT solutions into path classes for conformational transitions of alanine dipeptide that need to cross $\phi = 0°$ or $\phi = 120°$.

the other hand, the topography of the landscape viewed from P_{II}/C_5 is rather different. Although the potential energy of S_5 is slightly higher than for S_6, the valley leading from P_{II}/C_5 to the latter transition state is steeper, which means that energy variation associated to small moves is in average larger when moving from P_{II}/C_5 toward S_6 than when moving toward S_5. The steepness of these valleys, which mainly involves the variation of ψ, is approximately 0.03 kcal/mol per angular degree for $P_{II}/C_5 \rightarrow S_6$ and approximately 0.02 kcal/mol per degree for $P_{II}/C_5 \rightarrow S_5$. Since the T-RRT tree expands more favorably on easy slopes, transitions across S_5 are more probably found. In addition, the valley mounting from P_{II}/C_5 to S_5 is wider than the one from P_{II}/C_5 to S_6. The width of the former pathway is reflected by the more significant variability of class-I paths compared to class-II paths. As explained for the *Alien* benchmark, T-RRT has a higher probability to find paths through large passages than through narrow corridors. Since the width of a pathway is related with the entropy, we can argue that, for relatively similar variations of potential energy, T-RRT reaches more easily transition states through pathways that exhibit a larger entropic term.

Transition paths were also computed between pairs of minima $\{\alpha_R, C_7^{ax}\}$ and $\{C_5, C_7^{ax}\}$. The solutions are represented in Figures 9 and 10, respectively. Table 7 presents the distributions of paths classes. These transitions, which require to go across energy barriers around $\phi = 0°$ or $\phi = 120°$, are more complex than the two analyzed above, and a larger variety of transition path classes is propounded in related literature [44, 46, 47, 49]. Results obtained with T-RRT show the ability of the algorithm to capture such a variety of possible transition paths. Overall, the projection of the computed paths on the $\{\phi, \psi\}$ map shows a good fitting with the valleys of the energy landscape (remind that the figures shows a two-dimensional projection of the results, while the exploration takes place in a seven-dimensional space). The figures and the table highlight significant differences on the paths distributions between direct and reverse transitions for the two pairs of minima. Paths from α_R and C_5 (in the negative range of ϕ) toward C_7^{ax} (in the positive range of ϕ) have a higher probability to traverse the energy barriers at $\phi \approx 0°$ across S_2, while the reverse transitions $C_7^{ax} \rightarrow \alpha_R$ and $C_7^{ax} \rightarrow C_5$ go more frequently across S_1. Note that such a behavior has also been reported in related works [44, 47], where conformational transitions of alanine dipeptide are computed with variants of MD methods. For the transitions $\alpha_R \rightarrow C_7^{ax}$, class-II paths seem the most natural ones, since they only require crossing one transition state (S_2), and they are the shortest ones. The most probable alternative paths are class-I paths. Although the energy barriers along these paths are lower that for class-II paths, three transition states (S_6, S_1 and S_3) need to be traversed, instead of only one. The picture is different for the reverse $C_7^{ax} \rightarrow \alpha_R$ transition. Paths starting at C_7^{ax} have a higher probability to reach α_L across S_3 or S_4 than to reach directly α_R across S_2. We attribute this phenomenon to the local

shape of the landscape, as explained above for the case of the $P_{II} \rightarrow \alpha_R$ transition. Indeed, the three transition states have relatively similar energies, but the valleys leading from C_7^{ax} to S_3 or S_4 are much easier than the one leading to S_2. Note that the overall shape of the low-energy region around C_7^{ax} is significantly wider in the ψ direction than in the ϕ direction. Once in α_L, the most probable transition paths to α_R go across S_1 and S_5, or alternatively S_1 and S_6. Results for the $C_5 \leftrightarrow C_7^{ax}$ transitions are coherent with respect of those for $\alpha_R \leftrightarrow C_7^{ax}$. For the direct transitions, paths though S_2 are slightly more probable than paths though S_1, since the transition energy associated to pathway $C_5 \rightarrow S_5 \rightarrow \alpha_R \rightarrow S_2 \rightarrow C_7^{ax}$ is slightly lower than that of pathway $C_5/P_{II} \rightarrow S_1 \rightarrow \alpha_L \rightarrow S_3 \rightarrow C_7^{ax}$ (10.1 kcal/mol and 10.7 kcal/mol, respectively), and because the steepness and the width of the valleys mounting from C_5/P_{II} to S_1 and from α_R to S_2 is comparable. For the reverse transition, however, the great steepness of pathway $C_7^{ax} \rightarrow S_2$ compared to $C_7^{ax} \rightarrow S_3$ or $C_7^{ax} \rightarrow S_4$ seems to hinder transitions across S_2. Finally mention that, in all the cases, transitions across barriers at $\phi \approx 120^\circ$ are infrequent compared to transitions across $\phi \approx 0^\circ$. These infrequent pathways have also been obtained with other methods [44, 49].

4 Conclusion

We have proposed a novel method, called T-RRT, to explore conformational energy landscapes. The method combines recent path planning algorithms from the field of robotics with basic concepts of statistical physics. The T-RRT algorithm can be applied to find reachable energy minima from an arbitrary conformation. More interestingly, the same algorithm (possibly with a different parameter setting for improving performance) can also be applied to compute conformational transition paths between pairs of minima. T-RRT can be advantageous compared to many path sampling methods based on an initial trajectory, or requiring a reaction coordinate that biases the search. Simple benchmarks have been used in this work to validate the approach, and to facilitate the interpretation of the main features of the method. Results show that paths computed by T-RRT do not only depend on the potential energy of the explored states but are also affected by the overall shape of the energy landscape. Apparently, the exploration tends to favor the lowest free energy routes. However, it is difficult to quantify accurately the importance of energetic and entropic contributions within the conformational exploration. This would require further theoretical studies and a deeper analysis or results that remain for future work.

Finally, this paper aims to provide a basic algorithmic framework that could be extended for treating more complex systems. Similarly to MC-based methods, more sophisticated sampling schemes could be devised to enhance the efficiency of the exploration. Additionally, when a target conformation is specified, a biased scheme could be used to drive more quickly the tree toward the final conformation. In the short future, we expect to investigate extensions of T-RRT to yield a more efficient exploration of the conformational space of longer polypeptides.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation under project DPI2010-18449. Léonard Jaillet was supported by CSIC under JAE-Doc fellowship.

Footnotes

[1] These benchmarks were first proposed at the 2005 Workshop on Conformational Dynamics in Complex Systems.

References

[1] Liwo, A.; Czaplewski, C.; Scheraga, H. A. Curr. Opin. Struct. Biol. 2008, 18, 134.

[2] Christen, M.; van Gunsteren, W. F. J. Comp. Chem. 2008, 29, 157.

[3] Schön, J. C.; Jansen, M. Int. J. Mat. Res. 2009, 100, 135.

[4] Leach, A. Molecular Modelling: Principles and Applications, 2nd ed.; Prentice Hall, 2001.

[5] Frenkel, D.; Smit, B. Understanding Molecular Simulation. From Algorithms to Applications (Computational Science Series, Vol 1), 2nd ed.; Academic Press, Boca Raton, FL, 2001.

[6] Earl, D. J.; Deem, M. W. Phys. Chem. Chem. Phys. 2005, 7, 3910.

[7] Rauscher, S.; Neale, C.; Pomès, R. J. Chem. Theory Comput. 2009, 5, 2640.

[8] Torrie, G.M.; Valleu, J.P. J. Comput. Phys. 1977, 23, 187.

[9] Beutler, T.C., van Gunsteren, W. F. J. Chem. Phys. 1994, 100, 1492.

[10] Mousseau, N.; Derreumaux, P.; Barkema, G. T.; Malek, R. J. Mol. Graphics Modell. 2001, 19, 78.

[11] Laio. A; Piranello, M. Proc. Nat. Acad. Sci. USA 2002, 99, 12562.

[12] Bolhuis, P.G.; Dellago, C. Reviews of Computational Chemistry, Vol 27. Lipkowitz, K.B. (ed.). John Wiley and Sons, New York, N.Y. 2010

[13] Wales, D. J.; Scheraga, H. A. Science. 1999, 285, 1368.

[14] McGarrah, D. B.; Judson, R. S. J. Comput. Chem. 1993, 14, 1385.

[15] Wilson, S. R.; Cui, W.; Moskowitz, J. W.; Schmidt, K. E. J. Comp. Chem. 1991, 12, 342.

[16] Wales, D. J.; Doye, J. P. K. J. Phys. Chem. A 1997, 101, 5111.

[17] Ji, M.; Klinowski, J. Proc. R. Soc. A 2006, 462, 3613.

[18] Mills, G.; Jónsson, H. Phys. Rev. Lett. 1994, 72, 1124.

[19] Henkelman, G.; Uberuaga, B. P.; Jónsson, H. J. Chem. Phys. 2000, 113, 9901.

[20] E, W.; Ren, W.; Vanden-Eijnden, E. Phys. Rev. B 2002, 66, 052301.

[21] E, W.; Ren, W.; Vanden-Eijnden, E. J. Phys. Chem. B 2005, 109, 6688.

[22] Allen, R. J.; Valeriani, C.; ten Wolde, P. R. J. Phys.: Condens. Matter 2009, 21, 463102.

[23] Dellago, C.; Bolhuis, P.G.; Csajka, F.S.; Chandler, D. J. Chem. Phys. 1998, 108, 1964.

[24] Paci E.; Karplus, M. J. Mol. Biol. 1999, 288, 441.

[25] Schlitter, J.; Engels, J.; Krüger, P.; Jacoby, E.; Wollmer, A. Mol. Sim. 1993, 10, 291.

[26] Mouawad, L.; Perahia, D. J. Mol. Biol. 1996, 258, 393.

[27] Sibani, P.; van der Pas, R.; Schén, J. C. Comp. Phys. Comm. 1999, 116, 17.

[28] Schön, J. C.; Putz, H.; Jansen, M. J. Phys.: Condens. Matter 1996, 8, 143.

[29] Cortés, J.; Siméon, T.; Remaud-Siméon, M.; Tran, V. J. Comput. Chem. 2004, 25, 956.

[30] Yao, P.; Dhanik, A.; Marz, N.; Propper, R.; Kou, C.; Liu, G.; van den Bedem, H.; Latombe, J.-C.; Halperin-Landsberg, I.; Altman, R. B. IEEE/ACM Trans. Comput. Biol. Bioinform. 2008, 5, 534-545.

[31] Kirillova, S.; Cortés, J.; Stefaniu, A.; Siméon, T. Proteins 2008, 70, 131.

[32] Raveh, B.; Enosh, A.; Schueler-Furman, O.; Halperin, D. PLoS Comput. Biol. 2009, 5, e1000295.

[33] Chiang, T. H.; Apaydin, M. S.; Brutlag, D. L.; Hsu, D.; Latombe, J.-C. J. Comput. Biol. 2007, 14, 578.

[34] Amato, N. M.; Dill, K. A.; Song, G. J. Comput. Biol. 2003, 10, 239.

[35] Cortés, J.; Siméon, T.; Ruiz de Angulo, V.; Guieysse, D.; Remaud-Siméon, M.; Tran, V. Bioinformatics 2005, 21, 116.

[36] Cortés, J.; Le, D.T.; Iehl, R.; Siméon, T. Phys. Chem. Chem. Phys. 2010, 12, 8268.

[37] Jaillet, L.; Cortés, J.; Siméon, T. IEEE Trans. Robot. 2010, 6, 635.

[38] Battiti, R.; Brunato, M.; Mascia, F. Reactive Search and Intelligent Optimization. Operations Research/Computer Science Interfaces Series Vol. 45. Springer Verlag, Berlin. 2008.

[39] LaValle, S. M.; Kuffner, J. J. In: Algorithmic and Computational Robotics - New Directions, 2000; p 293.

[40] LaValle, S. M. Planning Algorithms; Cambridge University Press: New York, 2006.

[41] Belta, C.; Esposito, J.; Kim, J.; Kumar, V. Int. J. Robot. Res. 2005, 24, 219.

[42] Ettlin, A.; Bleuler, H. In: Proc. Int. Conf. on Control, Automation, Robotics and Vision. 2006, 1.

[43] Brooks III, C. L.; Case, D. A. Chem. Rev. 1993, 93, 2487.

[44] Apostolakis, J.; Ferrara, P.; Caflisch, A. J. Chem. Phys. 1999, 110, 2099.

[45] Bolhuis, P. G.; Dellago, C.; Chandler, D. Proc. Natl. Acad. Sci. USA 2000, 97, 5877.

[46] Chekmarev, D. S.; Ishida, T.; Levy, R. M. J. Phys. Chem. B 2004, 108, 19487.

[47] van der Vaart, A.; Karplus, M. J. Chem. Phys 2005, 122, 114903.

[48] Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. Multiscale Model. Simul. 2006, 5, 1214.

[49] Jang, H; Woolf, T. B. J. Comput. Chem. 2006, 27, 1136.

[50] Okumura, H.; Okamoto, Y. J. Phys. Chem. B 2008, 112, 12038.

[51] Strodel, B.; Wales, D. J. Chem. Phys. Lett. 2008, 466, 105.

[52] Velez-Vega, C.; Borrero, E. E.; Escobedo, F. A. J. Chem. Phys. 2009, 130, 225101.

[53] Liu, Z.; Ensing, B.; Moore, P.B. J. Chem. Theory Comput. 2011, 7, 402.

[54] Kollman, P. A.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. Comp. Simul. Biomol. Systems 1997, 3, 83.

[55] Corcho, F. J.; Filizola, M.; Pérez, J. J. Chem. Phys. Lett. 2000, 319, 65.

Figure Captions

Algorithm 1: Transition-based RRT.

Figure 1: RRT construction scheme. In blue/thin lines, the RRT tree. In red/bold lines, the Voronoi cells associated with the states contained in the tree. At each step, a state q_{rand} is randomly sampled, and its nearest neighbor in the search tree q_{near} is selected. It corresponds to the node in the Voronoi cell where q_{rand} has been sampled. A new node q_{new} is created by moving from q_{near} a distance δ in the direction of q_{rand}. The Voronoi bias favors the tree expansion toward unexplored regions of the space.

Figure 2: Impact of the Exploration Guarantee on the performance of T-RRT. The trees in both pictures have the same size (800 nodes), and are rooted at the same coordinate (-30, -30). Without this filtering process (a), the insertion of nodes very close to existing ones tends to slow down the exploration by decreasing the temperature. With filtering (b), the exploration of new regions of the space is favored.

Figure 3: *Zorro* potential. In bold black, a T-RRT transition path found between minima A and B. It circumvents two energy barriers and passes through a higher energy saddle-point. In thin black, other branches of the associated search tree.

Figure 4: *Alien* potential. Minima A and B can be connected through two pathways, u and l. In bold black, two paths found with T-RRT. In thin black, branches of the associated search trees.

Figure 5: Alanine dipeptide and the seven conformational parameters used for the exploration.

Figure 6: Energy minima and main saddle points found by T-RRT for the alanine dipeptide projected on the $\{\phi, \psi\}$ energy map.

Figure 7: Fifty representative paths for $\alpha_L \rightarrow C_7^{ax}$ (left) and $C_7^{ax} \rightarrow \alpha_L$ (right) transitions obtained with T-RRT.

Figure 8: Fifty representative paths for $\alpha_R \rightarrow P_{II}$ (left) and $P_{II} \rightarrow \alpha_R$ (right) transitions obtained with T-RRT.

Figure 9: Fifty representative paths for $\alpha_R \rightarrow C_7^{ax}$ (left) and $C_7^{ax} \rightarrow \alpha_R$ (right) transitions obtained with T-RRT.

Figure 10: Fifty representative paths for $C_5 \rightarrow C_7^{ax}$ (left) and $C_7^{ax} \rightarrow C_5$ (right) transitions obtained with T-RRT.

Algorithm 1: Transition-based RRT

input : the Conformational Space CS;
the energy function $E : CS \to \mathbb{R}_+^*$;
the initial conformation q_{init} ;
the target conformation q_{goal} (optional);
output : the tree $\mathcal{T}$;
begin
  $\mathcal{T} \leftarrow$ `InitTree`(q_{init});
  while not `StopCondition`$(\mathcal{T}, q_{goal})$ **do**
    $q_{rand} \leftarrow$ `SampleConf`(CS) ;
    $q_{near} \leftarrow$ `NearestNeighbor`$(q_{rand}, \mathcal{T})$;
    if `ExploGuarantee`$(\mathcal{T},\ q_{near},\ q_{rand})$ **then**
      $q_{new} \leftarrow$ `Extend`$(\mathcal{T}, q_{rand}, q_{near})$;
      if `TransitionTest`$(E(q_{near}), E(q_{new}))$ **then**
        `AddNewNode`$(\mathcal{T},\ q_{new})$;
        `AddNewEdge`$(\mathcal{T},\ q_{near},\ q_{new})$;
end

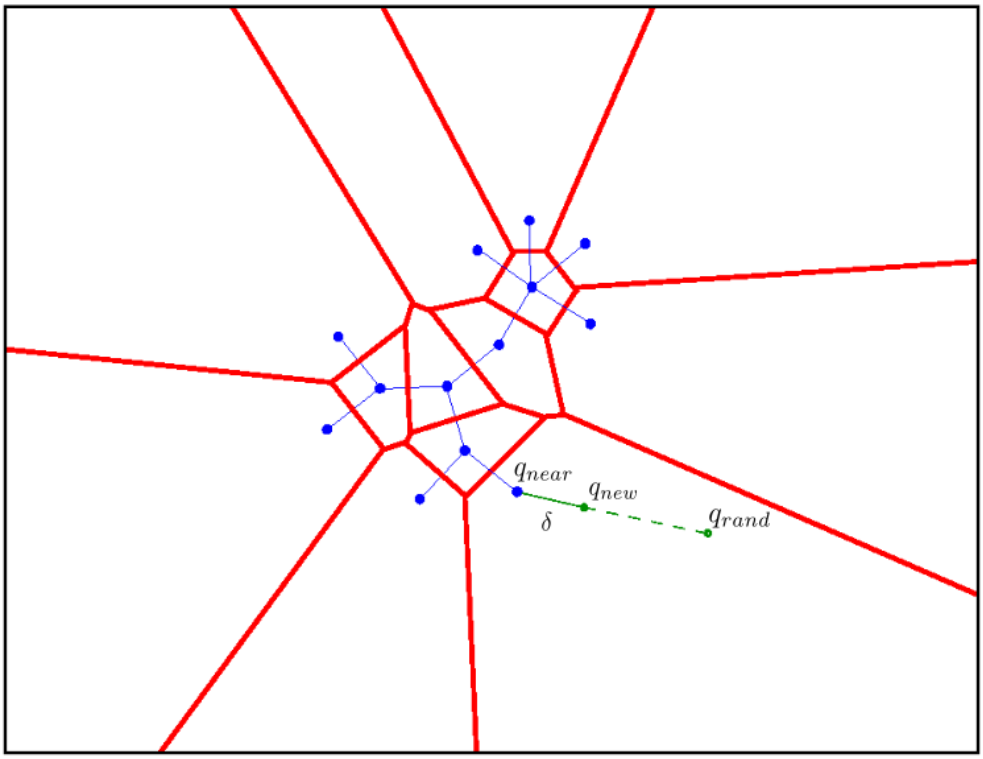

Figure 1:

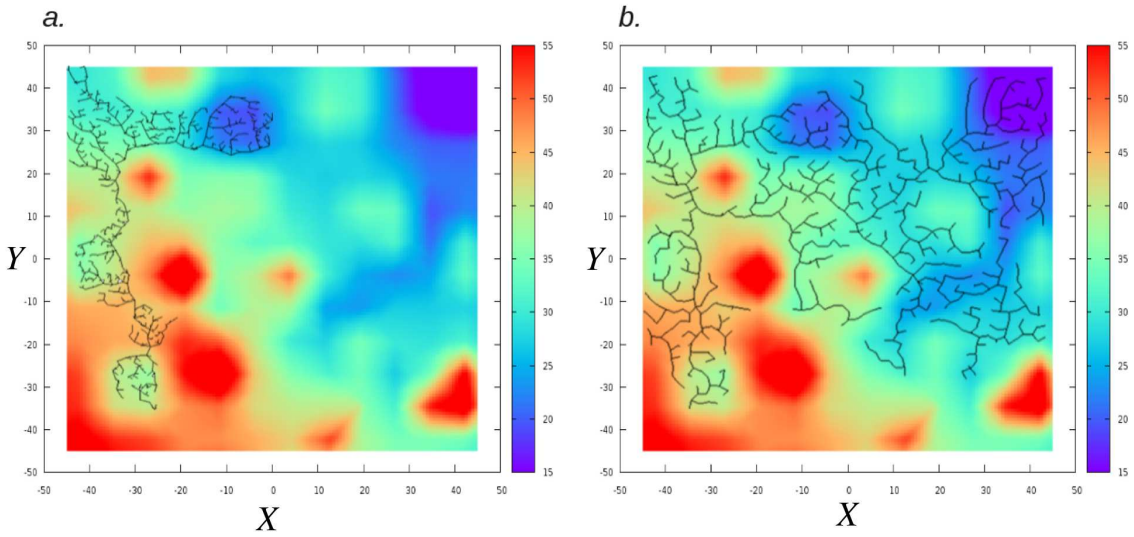

Figure 2:

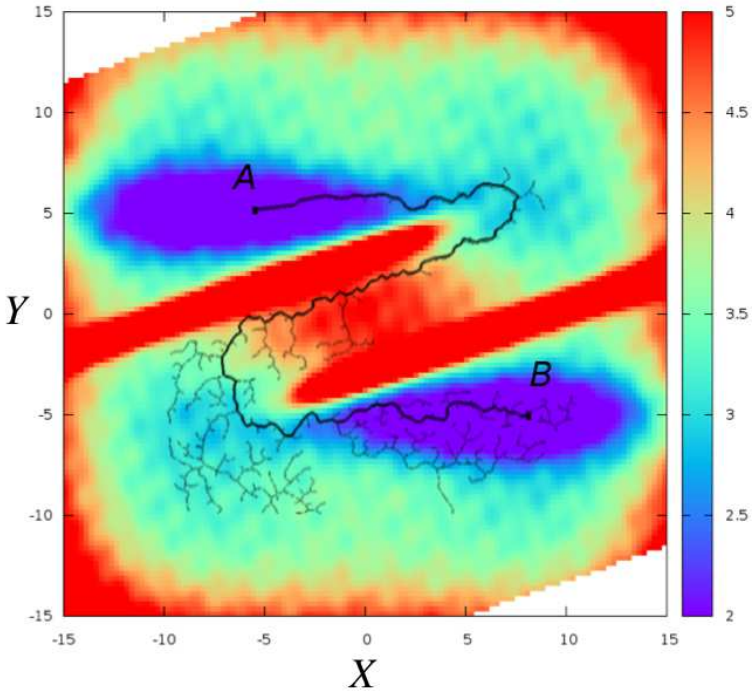

Figure 3:

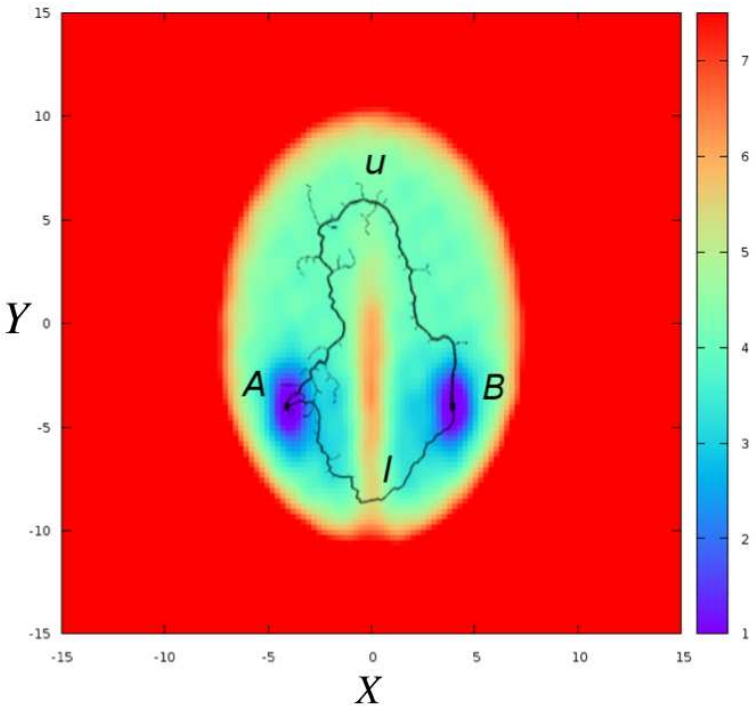

Figure 4:

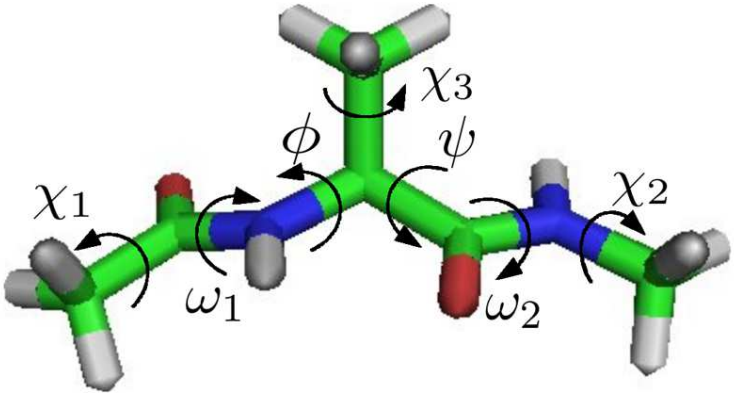

Figure 5:

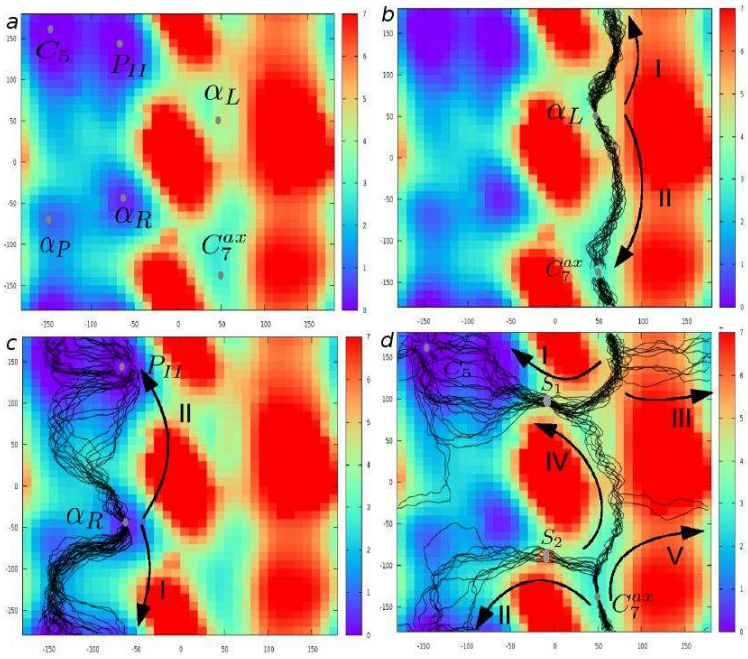
a
C_5
P_{II}
α_L
α_R
α_P
C_7^{ax}
b
α_L
I
II
C_7^{ax}
c
P_{II}
II
α_R
I
d
C_5
I
S_1
III
IV
S_2
V
II
C_7^{ax}

Figure 6:

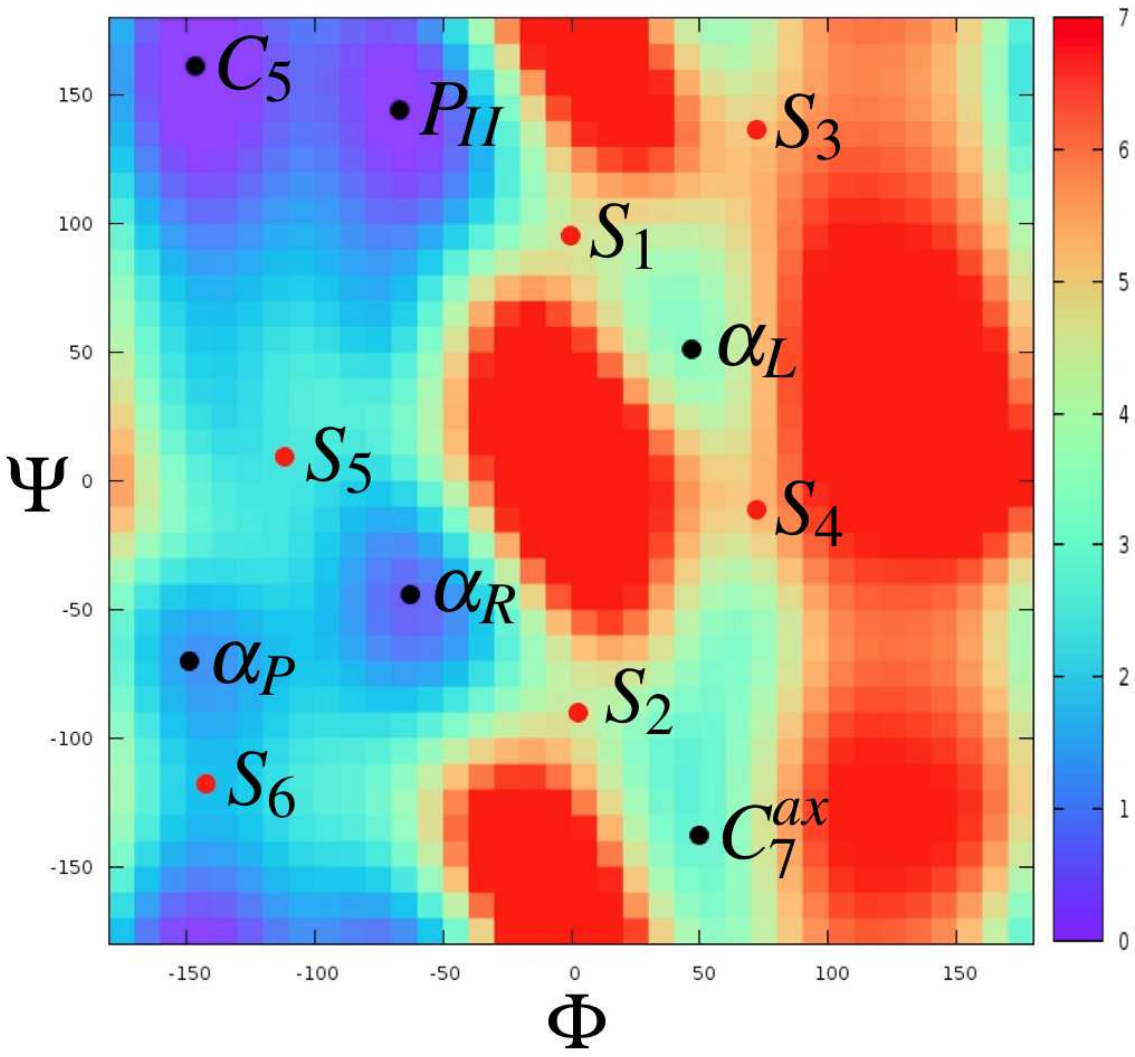
C_5
P_{II}
S_3
S_1
α_L
S_5
S_4
α_R
α_P
S_2
S_6
C_7^{ax}
150
100
50
0
-50
-100
-150
Ψ
-150
-100
-50
0
50
100
150
Φ
7
6
5
4
3
2
1
0

Figure 7:

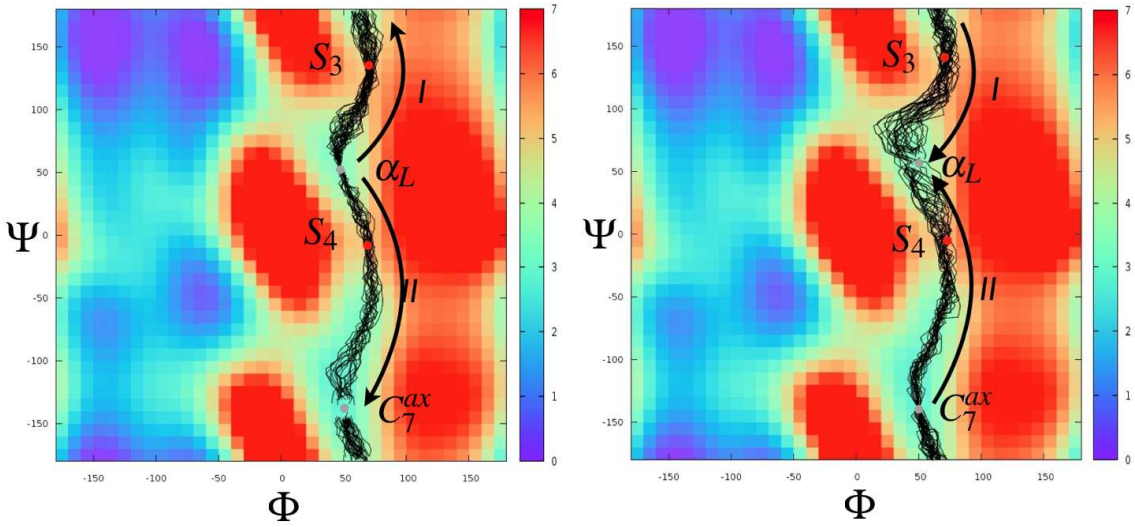

Figure 8:

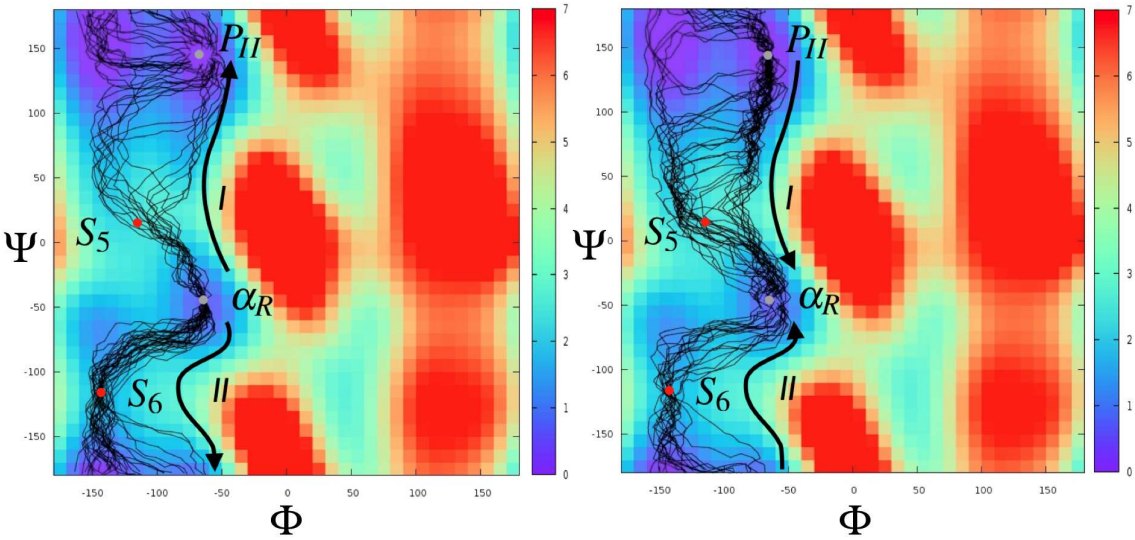

Figure 9:

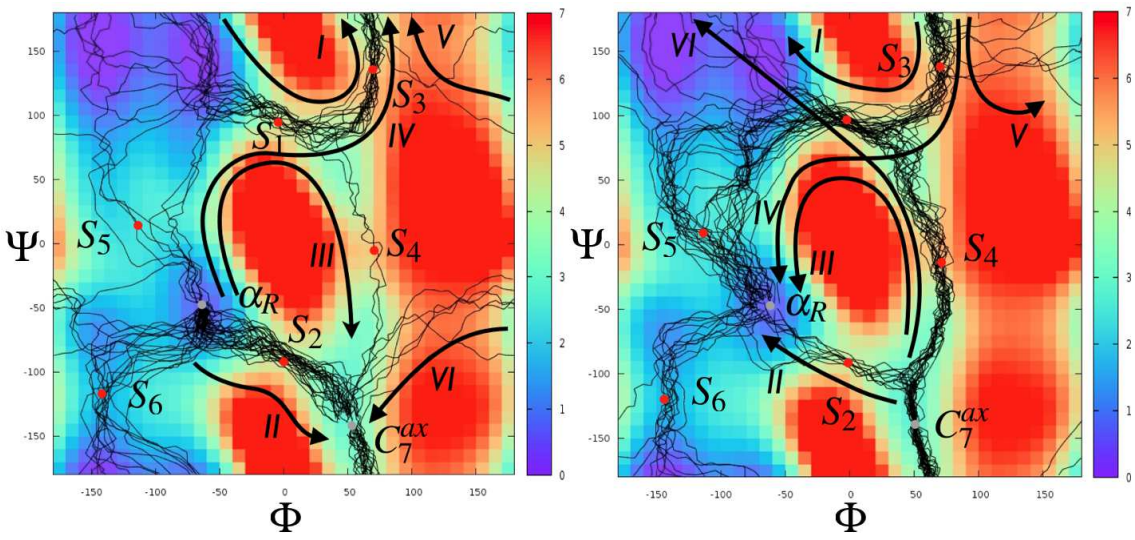
I
V
S_3
S_1
IV
S_5
III
S_4
α_R
S_2
VI
S_6
II
C_7^{ax}
VI
I
S_3
V
IV
S_5
III
S_4
α_R
II
S_6
S_2
C_7^{ax}
Ψ
Φ
Ψ
Φ

Figure 10: