A path planning approach for computing large-amplitude motions of flexible molecules

Juan Cortés, Thierry Simeon, Vicente Ruiz de Angulo, David Guieysse, Magali Remaud Simeon, Vinh Tran

A path planning approach for computing large-amplitude motions of flexible molecules

J. Cortés[1], T. Siméon[1,], V. Ruiz de Angulo[2], D. Guieysse[3], M. Remaud-Siméon[3] and V. Tran[4]*

[1]LAAS-CNRS, 7 av. du Colonel-Roche, 31077 Toulouse, France, [2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens Artigás 4-6, 2 planta, 08028 Barcelona, Spain, [3]Laboratoire Biotechnologie-Bioprocédés, UMR CNRS 5504, UMR INRA 792, INSA, 135 av. de Rangueil, 31077 Toulouse, France and [4]Unité Biotechnologie, Biocatalyse, Biorégulation (U3B), UMR CNRS 6204, Faculté des Sciences et Techniques, 2 rue de la Houssinière, 44322 Nantes, France

ABSTRACT

Motivation: Motion is inherent in molecular interactions. Molecular flexibility must be taken into account in order to develop accurate computational techniques for predicting interactions. Energy-based methods currently used in molecular modeling (i.e. molecular dynamics, Monte Carlo algorithms) are, in practice, only able to compute local motions while accounting for molecular flexibility. However, large-amplitude motions often occur in biological processes. We investigate the application of geometric path planning algorithms to compute such large motions in flexible molecular models. Our purpose is to exploit the efficacy of a geometric conformational search as a filtering stage before subsequent energy refinements.

Results: In this paper two kinds of large-amplitude motion are treated: protein loop conformational changes (involving protein backbone flexibility) and ligand trajectories to deep active sites in proteins (involving ligand and protein side-chain flexibility). First studies performed using our two-stage approach (geometric search followed by energy refinements) show that, compared to classical molecular modeling methods, quite similar results can be obtained with a performance gain of several orders of magnitude. Furthermore, our results also indicate that the geometric stage can provide highly valuable information to biologists.

Availability: The algorithms have been implemented in the general-purpose motion planning software Move3D, developed at LAAS-CNRS. We are currently working on an optimized stand-alone library that will be available to the scientific community.

Contact: nic@laas.fr

*To whom correspondence should be addressed.

1 INTRODUCTION

Interest in the computational analysis of molecular motions is well known. Macromolecular flexibility remains the main challenge for accurate docking approaches (Janin *et al*., 2003) or for studying molecular pathways (e.g. protein folding, structural rearrangements). Classical biomolecular modeling methods (Leach, 1996) are too computationally expensive for generating large motions while accounting for molecular flexibility. Molecular dynamics simulations are, in practice, applicable to computing motions in the time range of nanoseconds. Techniques based on Monte Carlo algorithms enable the computation of larger motions, but they are also constricted. In practice, the two limiting factors of these methods are the high cost for energy computation and their strong tendency to fall in the local minima of the energy landscape.

Motion planning algorithms (Latombe, 1991) originally developed in robotics are efficient tools for exploring constrained spaces. While these techniques have recently been extended to explore molecular force fields, our aim is to exploit the efficacy of a geometric treatment of molecular constraints in order to better handle the complexity of large amplitude motions and flexible molecular models. Such geometric treatment is already applied in several other works. For example, the geometric complementarity of molecular surfaces is a widely used criterion to predict protein–ligand or protein–protein interactions (Kuntz *et al*., 1982; Rarey *et al*., 1996; Jackson *et al*., 1998), especially in rigid docking approaches. Moreover, techniques for the conformational sampling of protein segments (Moult and James, 1986; DePristo *et al*., 2003; Lei *et al*., 2004) often apply geometry-based approaches to the loop closing problem and to atom overlap detection.

The driving idea of our approach is to separate the conformational search in two stages aiming to highly speed up the computation. The first stage consists in a geometric

filtering operated by motion planning techniques applied on articulated hard sphere models. The second stage accounts for the energy-based accuracy only for selected solutions found at the previous stage. The interest of this geometric filtering is that high-dimensional conformational spaces can be globally explored in a continuous way.

Section 2 overviews motion planning techniques and discusses their recent applications to computational biology. Section 3 describes the molecular models and Section 4 the tailored robotics algorithms developed for the geometric filtering stage of our approach. Classical molecular modeling techniques are currently used for the energy refinement stage. The approach is then applied to two general computational problems in biology. Section 5 deals with the analysis of protein loop mobility. Section 6 deals with the study of accessibility problems in protein–ligand interactions, considering both the flexibility of the ligand and that of the protein side-chains. Both studies show the efficacy of the approach for problems involving large-amplitude motions which are poorly treated by classical molecular modeling techniques. They also highlight the potential usefulness of the geometric treatment for guiding the rational design of proteins.

2 ROBOT MOTION PLANNING

Motion planning is a classical problem in robotics (Latombe, 1991). It consists in computing feasible motions for articulated robots in workspaces cluttered with obstacles. In recent years, motion planning techniques have undergone considerable development and have been successfully applied to challenging problems in diverse application domains, including computational biology.

2.1 Sampling-based motion planning algorithms

Sampling-based motion planners have been designed for exploring constrained high-dimensional spaces. Most of them are variants of the probabilistic roadmap (PRM) approach (Kavraki *et al*., 1996). The basic principle of PRM is to compute a connectivity graph (the roadmap) encoding representative feasible paths in the search space (e.g. the molecular conformational space). Nodes correspond to randomly sampled points that satisfy feasibility requirements (e.g. collision-freeness) and edges represent feasible subpaths computed between neighboring samples. Once the roadmap has been constructed, it is subsequently used to process multiple motion queries or to determine ensemble properties of mobility.

Variants of the PRM framework have been designed for solving single-query problems without preprocessing the complete roadmap. For example, the rapidly-exploring random trees (RRT) algorithm (LaValle and Kuffner, 2001) expands random trees rooted at the query positions and advancing towards each other through the use of a greedy heuristic. Such variants are well suited to highly constrained problems for which the solution space has the shape of a long thin tube. Whereas contructing a roadmap within the tube would require a high density of samples, the random tree variant benefits from the shape of the tube to naturally steer the expansion.

2.2 Applications to computational biology

Recently, PRM-based algorithms have been successfully applied to study molecular motions involved in biological processes such as protein–ligand interactions (Singh *et al*., 1999; Apaydin *et al*., 2004), protein folding (Amato *et al*., 2003; Apaydin *et al*., 2002) and also RNA folding (Tang *et al*., 2004). The main difference in the molecular adaptation of the PRM framework is that the binary collision detection, used in robotic applications, is replaced by a molecular force field. Sampled conformations are accepted on the basis of their potential energy and roadmap edges are weighted according to their energy cost. Although the framework is general enough to use any molecular force field, the techniques above generally consider simple potentials (including van der Waals and electrostatic terms) for the sake of efficiency.

The major strength of these sampling-based techniques is their ability to circumvent the energy trap problem encountered by classical simulation techniques, which waste a lot of time trying to escape from the local minima of the molecular energy landscape. Singh *et al*. (1999) and Apaydin *et al*. (2004) showed promising results from the study of binding sites for flexible ligands, assuming a rigid model of the protein to limit the dimension of the conformational space. Protein flexibility, which plays an important role in protein–ligand interactions however, is considered for protein folding applications using simplified models such as articulated backbone with bounding spheres for the side-chains (Amato *et al*., 2003) or a vector-based approximation of secondary structure elements (Apaydin *et al*., 2002).

Apart from our previous work on long protein loop conformational studies (Cortés *et al*., 2004), RRT-like methods have never been applied to computational biology problems.

3 GEOMETRIC MODELING

This section describes the geometric constraints considered by our approach to translate the driving forces affecting molecular motions. It then presents the molecular models handled by the motion planning algorithms.

3.1 Geometric view of molecular constraints

3.1.1 Molecular degrees of freedom Molecular mechanics force fields consider bonded and non-bonded atomic interactions separately. Bonded interactions concern the variation in the relative position of bonded atoms which is usually given in internal coordinates: bond lengths (stretching), bond angles (bending) and dihedral angles (torsion). Slight variations in bond lengths and bond angles from their ideal values produce a large increase in energy. Due to the stiffness of these two

terms, both parameters are generally kept constant and the molecular chain is considered as an articulated mechanism with revolute joints modeling bond torsions.

3.1.2 Loop closure constraints In many studies, the global molecular architecture is known and only segments of the molecular chain (loops and unstructured segments not involved in the secondary elements) are possible flexible elements. The first and last atoms of these flexible segments must remain bonded to the fixed neighboring atoms in the chain. Thus, kinematic loop closure constraints are introduced in the molecular chain. They reduce the subset of feasible conformations of the articulated molecular model. Similar constraints also appear in cyclic molecules and in the presence of disulphide bonds.

3.1.3 Main repulsive constraints For non-bonded interactions, the repulsive part of the van der Waals term is the most important contributor. A large amount of energy is required to get two non-bonded atoms significantly closer than the van der Waals equilibrium distance. Thus, acceptable conformations must respect geometric constraints for steric clash avoidance.

3.1.4 Main attractive constraints Conversely, two important attractive interactions are responsible for the globular shape of macromolecules and strongly participate to molecular docking: the hydrophobic interactions and the hydrogen bonds. They restrain the relative locations of the involved atoms and, therefore, imply geometric (distance/orientation) constraints on the molecular model.

3.2 Geometric molecular model

Our algorithms deal with all-atom models of molecules, which are considered as articulated mechanisms with atoms represented by spheres. Groups of rigidly bonded atoms form the bodies and the articulations between bodies correspond to bond torsions. A cartesian coordinate frame is attached to each group. The relative location of consecutive frames is defined by a homogeneous transformation matrix, which is a function of the rotation angle between them. We follow a method similar to that of Zhang and Kavraki (2002) to define such frames and matrices between rigid groups. Figure 1 shows the mechanical model for an amino acid residue.

Our modeling can also take advantage of a known secondary structure. In this case, the rigid secondary structure elements (alpha helices and beta sheets) are modeled as rigid groups of backbone atoms with articulated side-chains. Since secondary structure elements are fixed in the model, loop closure constraints are introduced in the in-between segment. Similar closure constraints can also be introduced in the model of a molecule to consider non-bonded interactions, such as hydrogen bonds, that impose the spatial proximity between some atoms of the protein.

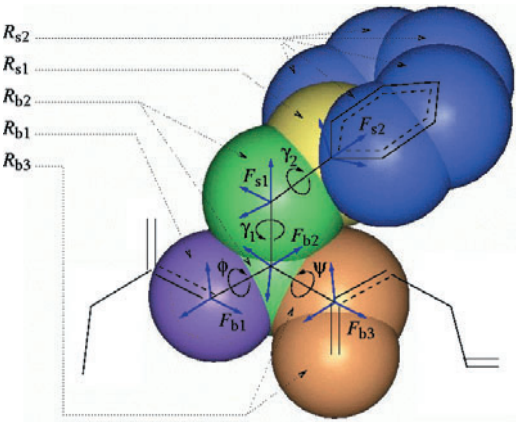

Fig. 1. Mechanical model of an amino acid residue (phenylalanin). It is composed of five rigid bodies, classified in:

— backbone rigids: $R_{b1} = \{N\}$, $R_{b2} = \{C_\alpha, C_\beta\}$, $R_{b3} = \{C, 0\}$;

— side-chain rigids: $R_{s1} = \{C_\gamma\}$, $R_{s2} = \{C_{\delta 1}, C_{\delta 2}, C_{\epsilon 1}, C_{\epsilon 2}, C_\zeta\}$.

The rotations between rigid atom groups are ϕ and ψ for the backbone, and γ_1 and γ_2 for the side-chain.

4 ALGORITHMS

In this section, we describe the motion planning algorithms developed for the geometric filtering stage of our approach. The conformational space is divided into feasible and forbidden regions. The feasible regions are defined as the subset of conformations avoiding steric clashes between the atoms of the articulated model, while satisfying the kinematic closure constraints associated with the loops and hydrogen bonds of the model. The main algorithm fulfills the conformational space exploration. Two principal functions called into this algorithm concern the sampling of points in the search space (i.e. conformational sampling) and the avoidance of steric clashes (i.e. collision detection).

4.1 Conformational space exploration

Conformational search is performed using a sampling-based motion planning technique. Molecular motions are in general extremely constrained mainly due to steric clashes and loop closure constraints. Therefore, we based our algorithm on RRT-like incremental search techniques (LaValle and Kuffner, 2001) that have been successfully applied to explore highly constrained spaces in other application domains such as mechanical disassembly.

The basic principle is to incrementally grow a random tree rooted at the initial conformation $\mathbf{q}_{\text{init}}$ to explore the reachable conformational space and find a feasible path connecting $\mathbf{q}_{\text{init}}$ to a goal conformation $\mathbf{q}_{\text{goal}}$. At each iteration, the tree is expanded toward a randomly sampled conformation $\mathbf{q}_{\text{rand}}$. This random sample is used to simultaneously determine the tree node to be expanded and the direction in which it is to be expanded (Fig. 2). The nearest node $\mathbf{q}_{\text{near}}$ in the tree to

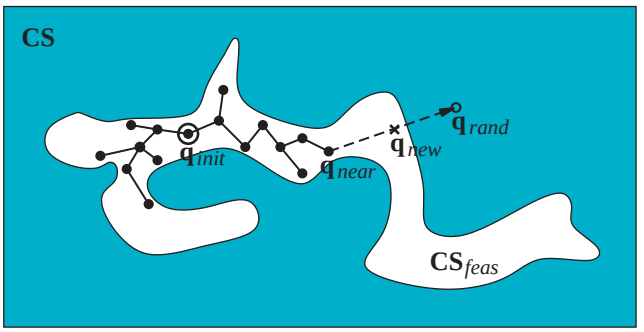

Fig. 2. Expansion of a search tree using an RRT-based algorithm.

Algorithm 1: Construct_RRT

```
input  : the root q_init, (optional) the goal q_goal
output : the tree τ, (optional) the path ρ
begin
    τ ← InitTree(q_init);
    while not StopCondition(τ, q_goal) do
        q_rand ← SampleConformation();
        q_near ← NearestNeighbor(τ, q_rand);
        q_new ← ExpandTree(q_near, q_rand);
        if not Similar(q_near, q_new) then
            AddNewNode(τ, q_new);
            AddNewEdge(τ, q_near, q_new);
    ρ ← ExtractPath(τ, q_init, q_goal);
end
```

the sample $\mathbf{q}_{\text{rand}}$ is selected and an attempt is made to expand $\mathbf{q}_{\text{near}}$ in the direction of the straight path to $\mathbf{q}_{\text{rand}}$. The key idea of this expansion strategy is to bias the exploration toward unexplored regions of the space. Hence, the probability that a node will be chosen for an expansion is proportional to the volume of its Voronoi region (i.e. the set of points closer to this node than to the others). Therefore RRTs are biased by large Voronoi regions to rapidly explore before uniformly covering the space.

Algorithm 1 gives the pseudocode for the RRT construction. The random conformations computed by the function `SampleConformation` can simply follow a uniform distribution over the space, as originally proposed for RRTs. We, however, prefer a more sophisticated sampling scheme recently introduced by Yershova *et al.* (2005). By limiting the uniform sampling inside domains dynamically computed around the explored regions, this variant was shown to outperform original RRTs on many constrained problems. Furthermore, the conformational sampler handles the presence of kinematic loop closure constraints (Cortés and Siméon, 2004) using the technique summarized in Section 4.2.

The function `ExpandTree` extracts the feasible portion of the path segment connecting $\mathbf{q}_{\text{near}}$ to $\mathbf{q}_{\text{rand}}$. The extremity $\mathbf{q}_{\text{new}}$ of the feasible subpath is computed by checking the satisfaction of the loop closure constraints and also the collision-freeness along the path segment. The overall performance of the RRT search strongly relies on the use of fast collision detection techniques. We use an efficient algorithm recently described by Ruiz de Angulo *et al.* (2005) and summarized in Section 4.3.

When the RRT search is performed to compute a feasible path connecting $\mathbf{q}_{\text{init}}$ to a goal conformation $\mathbf{q}_{\text{goal}}$, the expansion process is iterated until the current expanded node $\mathbf{q}_{\text{new}}$ can be connected to $\mathbf{q}_{\text{goal}}$, or a stop condition estimates that no solution exists. In the absence of a specified goal, the same algorithm can be used to encode within the computed tree a representative subset of feasible paths and conformations reachable from $\mathbf{q}_{\text{init}}$.

4.2 Loop conformational sampling

Molecular conformations satisfying loop-closure constraints are sampled based on the general technique called random loop generator (RLG) (Cortés *et al.*, 2002). Its application to protein loops is discussed by Cortés *et al.* (2004). RLG relies on a decomposition of the closed-chain mechanism. The kinematic chain corresponding to the loop backbone is divided into an active and a passive subchain. The passive subchain is a backbone portion involving six rotational bonds. The parameters (dihedral angles) of the active subchain are progressively sampled using a simple geometric algorithm that notably increases the probability of obtaining a conformation that satisfies loop closure. Once the active subchain has been sampled, the conformation of the passive one is computed by a general $6R$ inverse kinematics method (Renaud, 2000), which is an improved variant of the method proposed by Manocha and Canny (1994). RLG performs efficiently with long protein loops, which are a challenge for most other related techniques.

The loop conformational sampler also considers distance constraints between different elements of the closed chain. Such constraints enable to account for the presence of backbone hydrogen bonds, which particularly affect the motion of some loops, hairpin loops for example. Loop conformations are sampled such that the distance between N–O atom pairs involved in hydrogen bonds remains within a given range.

The algorithm optionally integrates collision detection into the progressive sampling process. Each time a dihedral angle is generated, overlaps between atoms in the corresponding rigid group and the rest of the protein are checked. The value is only kept if no collision exists. Conformations computed in this way simultaneously satisfy loop-closure and steric clash avoidance. This combined procedure is more efficient than that consisting of closing the loop first and then checking for collisions.

4.3 Collision detection

Performance requirement of collision detection within motion planning algorithms is especially important for molecular applications because of the quadratic cost of enumerating all non-bonded atom pairs in models with thousands of atoms. For this purpose, we have developed a tailored algorithm BioCD (Ruiz de Angulo *et al*., 2005) for efficient self-collision and distance computations in highly articulated molecular models. BioCD uses, like Lotan *et al*. (2002), hierarchical data structures that approximate the shape of the protein at successive levels of details, allowing the number of interacting pairs tested for collision to be significantly reduced. However, while the former algorithm was designed for Monte Carlo searches which only slightly change at each step a few randomly selected degrees of freedom (DOF), BioCD is more adapted to our sampling-based motion planning scheme in which much larger sets of DOF are simultaneously and arbitrarily modified during conformational space exploration.

BioCD is inspired by the dual kd-tree traversal algorithms initially developed for n-point correlation problems in statistical learning (Moore *et al*., 2001). The algorithm maintains two levels of bounding volume hierarchies grouped according to spatial proximity. The first level organizes the rigid parts of the articulated model according to the selected DOF while the second level organizes the atoms inside each rigid part of the first level. Such a data-structure can be efficiently tested for collision and also updated at a moderate cost. Experimental tests performed with BioCD show its efficacy in processing thousands of collision tests per second on articulated protein chains with hundreds of DOF (Ruiz de Angulo *et al*., 2005).

The next sections discuss two applications of this planning technique to problems involving large molecular motions. They also demonstrate the ability of the approach to handle efficiently articulated models with many DOF.

5 PROTEIN LOOP MOBILITY

Loops are irregular portions of proteins. Such 'irregularity' makes structure prediction difficult and therefore, currently available techniques often fail when applied to long loops (Tramontano *et al*., 2001). Indeed, surface loops can, in many cases, undergo significant conformational changes and adopt a variety of energetically favorable conformations. The main interest in studying loop conformational changes is due to their importance in protein interactions. For instance, they can adapt the surface topology of antibodies for antigen recognition (James *et al*., 2003) and play a key role in catalytic mechanisms (Osborne *et al*., 2001). Despite the importance of protein loops, very limited tools are available to analyze their mobility. Energy-based approaches are only applicable (in practice) to the computation of slight conformational changes. For larger motions, simpler computational approaches have to be designed. As far as we know, only the recent technique ROCK (Lei *et al*., 2004) is able to generate such loop motions in reasonable computing time. The results presented below show the good performance of our algorithm for studying the mobility of protein loops.

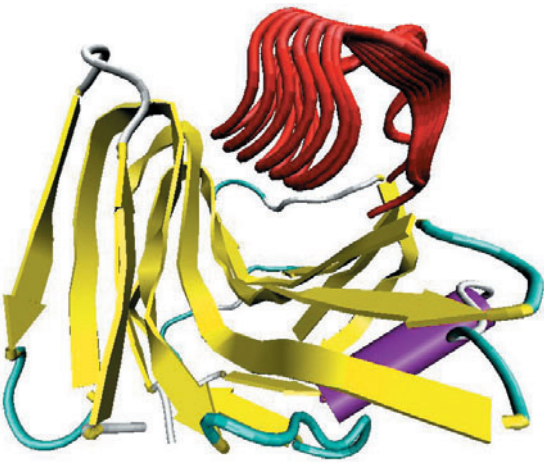

Fig. 3. Geometrically feasible conformational change of the 'thumb' hairpin loop of xylanase from *Thermobacillus xylanilyticus*.

5.1 Mobility of a specific loop of xylanase

We studied endo-β-1,4-xylanase (EC 3.2.1.8) from *Thermobacillus xylanilyticus* (XTX) aiming to optimize the conversion of cereal co-products into bio-ethanol fuel. The architecture of XTX[1] is similar to a right hand, where the thumb is a long hairpin loop (Fig. 3). In this protein, the corresponding amino acid sequence spans from 107 to 125. Although maintained by a network of hydrogen bonds, this loop is suspected to be very flexible, like in other xylanases (Muilu *et al*., 1998). An open loop conformation may allow an easier access of the substrate (xylan) to the catalytic pocket. Once the xylan is inside the main crevice, a closed conformation of this loop could complete the full docking.

In a previous molecular modeling study of XTX loop, a modified simulated annealing procedure was used to sample conformations while considering hydrogen-bond networks (HBN) that maintain the hairpin structure of the loop. The procedure can be summarized as follows. The initial dynamics simulations were performed at 500 K with a CFF91 force field from Accelrys, maintaining constraint distances (around 3.2 Å) between non-hydrogen atoms supposed to be involved in hydrogen bonds. Then, sets of conformations chosen regularly along the high temperature trajectories were progressively and slowly cooled and minimized ($>10\,000$ iterations). At the final minimization step, the distance constraints were removed. Only low energy conformations were selected and then clustered in several (10) low energy regions. Several possible HBN (from known structures of similar anti-parallel

[1]Known from personal communication. Structures of other xylanases of the same family are available (de Lemos *et al*., 2004).

beta-sheet loops) were tested. Results showed that significantly different low energy conformations of the 'thumb' loop are possible. However, the method has prominent drawbacks. Firstly, it is computationally too expensive. A complete set of calculations (dynamics + minimization) for a given HBN needed >10 h on an SGI computer with MIPS R14000 processor. Consequently, the whole study required several days. Besides, the method is unable to determine whether a continuous loop motion exists between two different feasible conformations.

5.2 Results

5.2.1 Conformational sampling Our first experiments with XTX aimed to validate our approach in relation to the classical molecular modeling method commented upon above. The considered articulated loop model involves 68 DOF (42 for the backbone and 26 for the side-chains). Sets of conformations satisfying loop closure and steric clash avoidance were computed for the different HBN by the conformational sampling algorithm (Section 4.2). Then, a simple energy minimization was performed. The method is very fast: 10 significantly different conformations are generated in some seconds or at most in some minutes, depending on the difficulty of satisfying a given HBN (on a PC with an Intel Pentium M 1.8 MHz processor). Since geometrically feasible conformations already satisfy some strong constraints, their energy minimization is also very fast, requiring about a minute per conformation. As shown in Figure 4, this minimization only produces a slight deformation of the loop conformation (RMSD < 1 Å). Notably, the energy values are in general very similar to those obtained by the classical modeling method. The whole computations takes <1 h instead of several days required without the geometric filtering stage.

5.2.2 Loop pathways and mobility analysis Next, we applied the motion planning algorithm (Section 4.1) to model the conformational change continuity, yielding a more accurate study of xylanase loop mobility. The tests showed the feasibility of a continuous loop motion from the crystallographic conformation to a conformation inside the crevice. The geometric path illustrated in Figure 3 was computed in <1 min.

In a further study, we applied the exploration algorithm to compare the mobility of the thumb loop in native XTX and in a mutant with two deletions: Tyr-111 and Thr-121. Experimentally, this mutant presents no activity. Tests were carried out with different HBN. In all cases, the mutated loop presented a much more restricted mobility in the crevice compared to the native one. Figure 5 shows the possible loop motions computed for the HBN represented in Figure 6. The two search trees (RRT) contain 5000 nodes and their construction required <1 h on a standard PC. These tests tend to confirm the impossibility for this mutated loop to go deep inside the crevice to fix the ligand for the catalytic action. This could explain the absence of activity experimentally observed for this mutant.

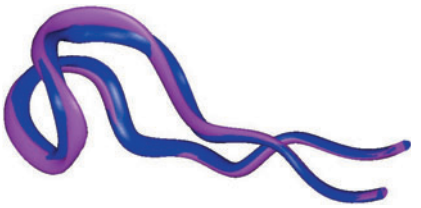

Fig. 4. Geometrically feasible random conformation of the xylanase loop (blue/dark), and low energy conformation obtained from it by simple minimization (pink/clear). Both conformations are very close, the backbone RMSD is 0.67.

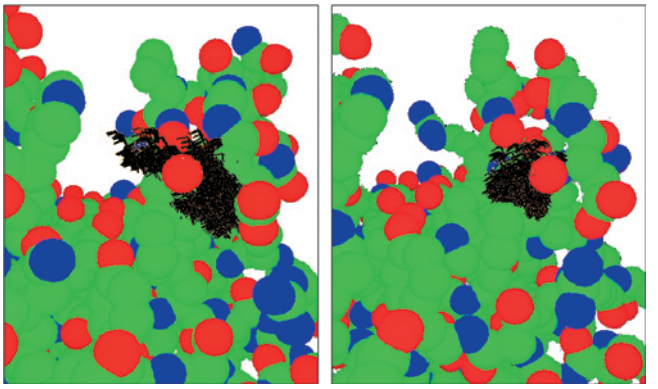

Fig. 5. Graphic representation of the loop mobility computed for native (left) and mutated xylanase (right). The small frames (in black) display positions reachable by the Cα atom of the middle residue in the loop. The native loop can move toward the crevice while the mutated loop only undergoes slight conformational changes.

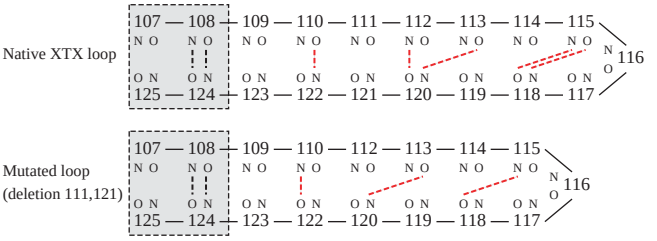

Fig. 6. HBN for the native/mutated loop of XTX. The HBN help to maintain the hairpin-like loop structure.

6 PROTEIN–LIGAND ACCESSIBILITY

Today, integrating protein flexibility in receptor-ligand interactions remains a challenge for accurate computer-aided drug design. Besides, the recent methods proposed for flexible docking (Carlson, 2002) mostly address the local aspect of the problem and compute the binding conformation of the ligand without considering the access pathway. The active

site of many enzymes is located at the bottom of a deep narrow cavity. In such cases, it is reasonable to consider that the docking of the ligand to the binding pocket is influenced by the difficulty of accessing the active site, which affects the enzyme–ligand affinity. Computing motions as large as that of a ligand entering from the protein surface to a deep active site remains computationally very expensive for energy-based modeling methods. We present below the results obtained with the proposed two-stage approach for studying the enantioselectivity of an enzyme presenting such a deep catalytic pocket.

6.1 Study of lipase enantioselectivity

The active site of the *Burkholderia cepacia* lipase (BCL) is located at the bottom of a narrow 17 Å deep pocket. This lipase is used for the kinetic resolution of racemic 2-substituted carboxylic acid in a transesterification reaction. Recently, a classical approach involving the modeling of the *R* and *S* tetrahedral intermediates of the reaction failed to explain the enzyme enantioselectivity (Guieysse *et al*., 2003). Thus, a molecular modeling procedure based on pseudo-molecular dynamics simulation under constraints was employed to model the trajectory of each enantiomer from the active site to the protein surface. Figure 7 shows the trace of the trajectory of one of the enantiomers. Interestingly, the energy of the enzyme/enantiomer interaction along the trajectory was found to be always lower for the preferred enantiomer which is in agreement with the experimental results of kinetic resolution. Consequently, molecular modeling of each enantiomer trajectory may be very helpful in understanding the enzyme enantioselectivity and very useful for its prediction. However, the modeling protocol designed by Guieysse *et al*. (2003) has several drawbacks. First, it does not enable automatic modeling because the manual correction of several side-chain orientations is required to remove the steric conflicts between the protein and the ligand along the trajectory. Besides, it is very time-consuming. Several days are required to generate one trajectory.

6.2 Results

6.2.1 Ligand pathways Our incremental search planner was used to compute geometrically feasible paths of articulated (*R*,*S*)-enantiomers, while also considering the flexibility of 17 side-chains in the catalytic pocket of BCL. The model contains a total of 68 DOF (11 for the ligand and 57 for the protein side-chains). Paths were computed for several couples of (*R*,*S*)-2-halogenophenyl acetic acid ethyl ester (referred to as ph(X)Et) in BCL for which both experimental (*in vivo*) results and *in silico* predictions are available. Computing times ranged from seconds to several minutes for solving the most constrained problems (Fig. 9). The first conclusion of the study is that all the ligand paths computed by our geometric approach are very similar to those obtained, after several days of computation, by pseudo-molecular dynamics.

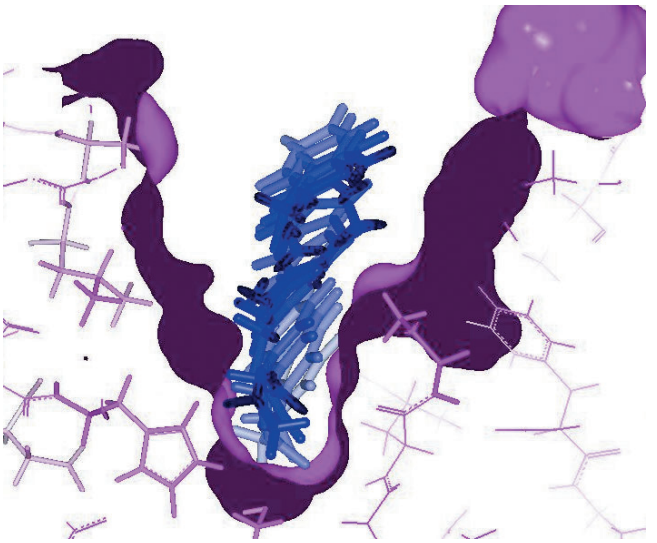

Fig. 7. Trajectory of a ligand ((*R*)-ph(Br)Et) accessing the active site of *Burkholderia cepacia* lipase.

After a simple and fast energy minimization of intermediate conformations along the geometrically feasible paths, the energy profiles are also very similar to the curves computed by pseudo-molecular dynamics (Fig. 8).

6.2.2 Enzyme enantioselectivity We applied the motion planner to compute geometric paths to the active site of BCL for several enantiomer pairs. Significantly different computing times were observed for some of them. Figure 9 shows the average times calculated over 50 runs of the planner. In general, the computing time of sampling-based motion planning algorithms increases with the difficulty of the problem. Thus, assuming that the topology of the catalytic pocket is better adapted to the access of the preferred enantiomer, its path should be computed faster than the path of the slow reacting one. Notably, results show a good correlation between the ratio of the computing time necessary to find the path of (*R*,*S*)-enantiomers and the experimental enantioselectivity (Guieysse *et al*., 2003). Therefore, these results indicate that the time spent by the planner may be useful information for predicting the enzyme enantioselectivity.

6.2.3 Mutagenesis targets In addition, an analysis of the set of computed paths enables a rapid localization of amino acid residues constraining the access of the enantiomers and is involved in the BCL discrimination of racemic compounds. The histograms in Figure 10 display the enzyme atoms constraining the motion of the (*R*,*S*)-ph(Br,F)Et enantiomers in four different portions of the path. It can be seen that the (*S*)-ph(Br)Et enantiomer meets a higher number of atoms restraining its access. This is totally consistent with computing time results. Therefore, this fast technique makes it very

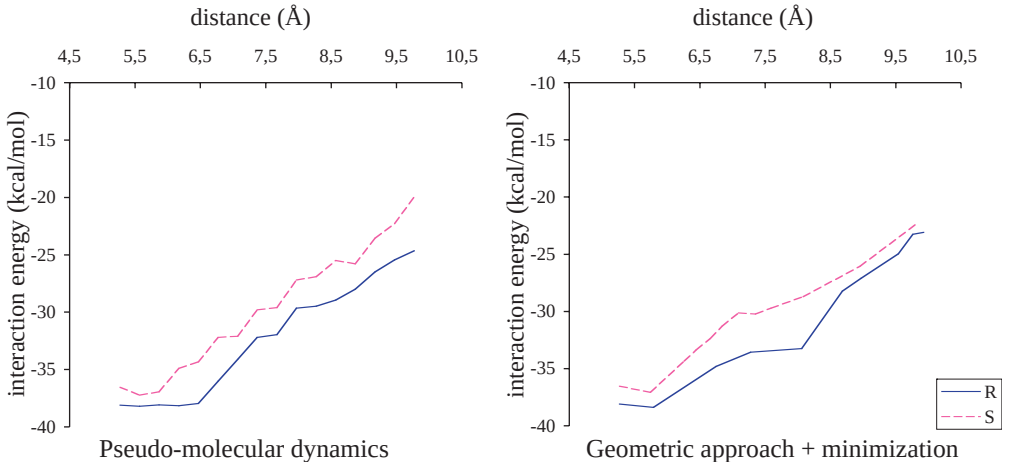

Fig. 8. Interaction energy along the most constrained portion of the ligand trajectory from the BCL active site. The distance is measured between the barycenters of the catalytic amino acid and the ligand. The curves obtained by pseudo-molecular dynamics (left) are very similar to those obtained by geometric motion planning followed by energy minimization (right). In both cases, the (*S*)-enantiomer displays a higher interaction energy.

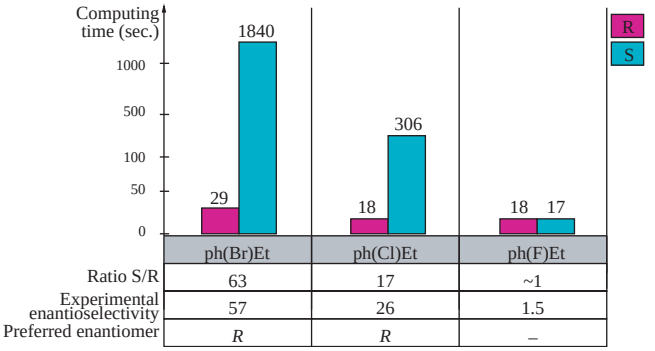

Fig. 9. Diagram representing the average time for computing paths for three pairs of enantiomers, (*R*,*S*)-ph(Br,Cl,F)Et. The computing time ratio *S*/*R* correlates with experimental enantioselectivity.

easy to pinpoint those residues possibly involved in enantioselectivity, thereby providing highly valuable information for site-directed mutagenesis.

7 DISCUSSION

The basic idea behind our approach is that major constraints affecting molecular motions have a geometric interpretation; this can be properly and rapidly dealt with by adequate tools such as robotics motion planning algorithms. The approach applies sampling-based motion planners to flexible molecular models as efficient conformational filters before an energy refinement. This fast geometric filtering computes energetically reasonable conformations and pathways allowing us to drastically accelerate the highly expensive computations of classical energy-based methods applied for the refinement stage. The resulting performance gain is particularly important to address problems involving large-amplitude motions in high-dimensional spaces, for which the applicability of energy-based approaches is limited. Furthermore, as shown by the results above, a biological interpretation can be directly made from geometrically feasible molecular motions.

Two kinds of application have been presented. The first one concerns the analysis of protein flexibility aimed at predicting possible conformational changes in polypeptide segments. The other application concerns problems of accessibility in protein–ligand interactions. Although the later study only involved protein side-chain flexibility, backbone flexibility (treated in the first application) can also be considered by the current algorithm.

Our geometric algorithms have been tested with several molecules and the first results are highly promising. However, an in-depth study with a larger set of models remains to be carried out. Currently, we are considering other examples of protein–ligand interactions involving deep narrow cavities and with available experimental results for a more rigorous validation.

This work focused on the geometric level of the approach while classical energy-based molecular modeling methods were used for the second stage. Our aim is to develop new techniques for this stage to better exploit the geometric path information provided in the first filtering stage.

ACKNOWLEDGEMENTS

This work has been partially supported by the PIR CNRS project BioMove3D, the European project IST-37185 MOVIE and the French ADEME project 02-01051.

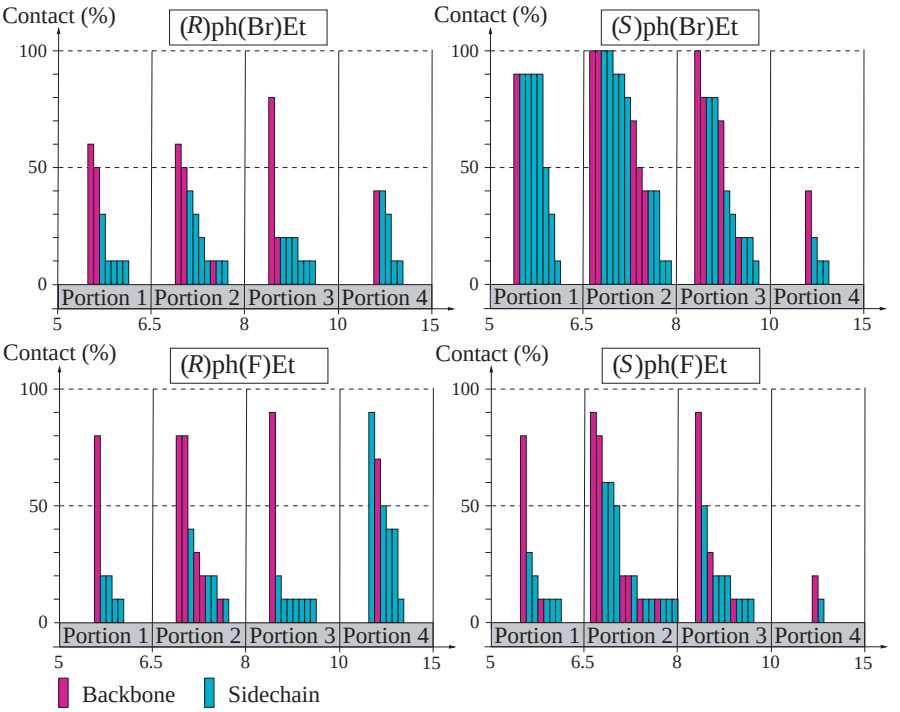

Fig. 10. Histograms representing lipase atoms constraining the ligand motion in different portions of the path. Contact is considered between 80 and 100% of the van der Waals equilibrium distance. The ordinates represent the percentage of times (over 50 runs) that a contact is observed. The names of the atoms and the corresponding residues have been omitted for reasons of clarity.

REFERENCES

Amato,N.M., Dill,K.A. and Song,G. (2003) Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, **10**, 149–168.

Apaydin,M.S., Brutlag,D.L., Guestrin,C., Hsu,D. and Latombe,J.-C. (2002) Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. In *Proceedings of RECOMB*, Washington, DC, pp. 12–21.

Apaydin,M.S., Brutlag,D.L., Guestrin,C., Hsu,D. and Latombe,J.-C. (2004) Stochastic conformational roadmaps for computing ensemble properties of molecular motion. In *Algorithmic Foundations of Robotics V (WAFR'02)*, Springer-Verlag, Berlin, pp. 131–147.

Carlson,H.A. (2002) Protein flexibility and drug design: how to hit a moving target. *Cur. Opin. Chem. Biol.*, **6**, 447–452.

Cortés,J., Siméon,T. and Laumond,J.-P. (2002) A random loop generator for planning the motions of closed kinematic chains using PRM methods. In *Proceedings of IEEE International Conference on Robotics and Automation*, Washington, DC, pp. 2141–2146.

Cortés,J., Siméon,T., Remaud-Siméon,M. and Tran,V. (2004) Geometric algorithms for the conformational analysis of long protein loops. *J. Comput. Chem.*, **25**, 956–967.

Cortés,J. and Siméon,T. (2004) Sampling-based motion planning under kinematic loop-closure constraints. In *Algorithmic Foundations of Robotics VI (WAFR'04)*, Springer-Verlag, Berlin. (in press).

DePristo,M.A., de Bakker,P.I.W., Lovell,S.C. and Blundell,T.L. (2003) *Ab initio* construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins*, **51**, 41–55.

Guieysse,D., Salagnad,C., Monsan,P., Remaud-Simeon,M. and Tran,V. (2003) Towards a novel explanation of *Pseudomonas Cepacia* lipase enantioselectivity via molecular modelling of the enantiomer trajectory into the active site. *Tetrahedron: Asymmetry*, **14**, 1807–1817.

Jackson,R.M., Gabb,H.A. and Stenberg,M.J.E. (1998) Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, **276**, 265–285.

James,L.C., Roversi,P. and Tawfik,D.S. (2003) Antibody multispecificity mediated by conformational diversity, *Science*, **299**, 1362–1367.

Janin,J., Henrick,K., Moult,J., Eyck,L.T., Sternberg,M.J.E., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: a critical assessment of predicted interactions, *Proteins*, **52**, 2–9.

Kavraki,L.E., Svestka,P., Latombe,J.-C. and Overmars,M.H. (1996) Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robotics Autom.*, **12**, 566–580.

Kuntz,I.D., Blaney,J.M., Oatley,S.J., Langridge,R. and Ferrin,T.E. (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **161**, 269–288.

Latombe,J.-C. (1991) *Robot Motion Planning*. Kluwer Academic Publishers, Boston, MA.

LaValle,S.M. and Kuffner,J.J. (2001) Rapidly-exploring random trees: progress and prospects. *Algorithmic and Computational Robotics: New Directions (WAFR2000)*, A.K. Peters, Boston, pp. 293–308.

Leach,A.R. (1996) *Molecular Modeling: Principles and Applications*. Longman, Essex.

Lei,M., Kuhn,L.A., Zavodszky,M.I. and Thorpe,M.F. (2004) Sampling protein conformations and pathways. *J. Comput. Chem.*, **25**, 1133–1148.

de Lemos,E.F., Esteves,F., Ruelle,V., Lamotte-Brasseur,J., Quinting,B. and Frère,J.M. (2004) Acidophilic adaptation of family 11 endo-β-1,4-xylanases: modeling and mutational analysis. *Protein Sci.*, **13**, 1209–1218.

Lotan,I., Schwarzer,F., Halperin,D. and Latombe,J.-C. (2002) Efficient maintenance and self-collision testing for kinematic chains. In *Proceedings of the 18th ACM Symposium on Computational Geometry*, Barcelona, Spain, pp. 43–52.

Manocha,D. and Canny,J. (1994) Efficient inverse kinematics of general 6R manipulators. *IEEE Trans. Robotics Autom.*, **10**, 648–657.

Moore,A.W., Connolly,A., Genovese,C., Gray,A., Grone,L., Kanidoris,N.,II, Nichol,R., Schneider,J., Szalay,A., Szapudi,I. and Wasseerman,L. (2001) Fast algorithms and efficient statistics: N-point correlation Functions. In *Proceedings of MPA/MPE/ESO Conference on Mining the Sky*, Springer-Verlag, New York.

Moult,J. and James,M.N.G. (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins*, **1**, 146–163.

Muilu,J., Torronen,A., Perakyla,M. and Rouvinen,J. (1998) Functional conformational changes of endo-1,4-xylanase II from *Trichoderma reesei*: a molecular dynamics study. *Proteins*, **31**, 434–44.

Osborne,M.J., Schnell,J., Benkovic,S.J., Dyson,H.J. and Wright,P.E. (2001) Backbone dynamics in dihdydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. *Biochemistry*, **40**, 9846–9859.

Rarey,M., Kramer,B., Lengauer,T. and Klebe,G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261**, 470–489.

Renaud,M. (2000) A simplified inverse kinematic model calculation method for all 6R type manipulators. In *Proceedings of International Conference on Mechanical Design and Production*, Cairo, Egypt, pp. 15–25.

Ruiz de Angulo,V., Cortés,J. and Siméon,T. (2005) BioCD: an efficient algorithm for self-detection and distance computation between highly articulated molecular models. In *Proceedings of Robotics: Science and Systems*, Cambridge, MA.

Singh,A.P., Latombe,J.-C. and Brutlag,D.L. (1999) A motion planning approach to flexible ligand binding. In *Proceedings of ISMB*, AAAI Press, Menlo Park, CA, pp. 252–261.

Tang,X., Kirkpatrick,B., Thomas,S., Song,G. and Amato,N.M. (2004) Using motion planning to study RNA folding kinetics. In *Proceedings of RECOMB*, Bertinoro, Italy, pp. 252–261.

Tramontano,A., Leplace,R. and Morea,V. (2001) Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, **Suppl. 5**, 22–38.

Yershova,A., Jaillet,L., Siméon,T. and LaValle,S.M. (2005) Dynamic-domain RRTs: efficient exploration by controlling the sampling domain. In *Proc. IEEE Int. Conf. Robotics Automation*, Barcelona, Spain, pp. 3867–3872.

Zhang,M. and Kavraki,L.E. (2002) A New Method for Fast and Accurate Derivation of Molecular Conformations. *J. Chem. Inf. Comput. Sci.*, **41**, 64–70.