Encoding Molecular Motions in Voxel Maps

Juan Cortés, Sophie Barbe, Monique Erard, Thierry Simeon

To cite this version:

Juan Cortés, Sophie Barbe, Monique Erard, Thierry Simeon. Encoding Molecular Motions in Voxel Maps. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8 (2), pp.557-563. 10.1109/TCBB.2010.23 . hal-02652128v2

HAL Id: hal-02652128
https://laas.hal.science/hal-02652128v2

Submitted on 18 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Encoding Molecular Motions in Voxel Maps

Juan Cortés, Sophie Barbe, Monique Erard and Thierry Siméon

Abstract—This paper builds on the combination of robotic path planning algorithms and molecular modeling methods for computing large-amplitude molecular motions, and introduces voxel maps as a computational tool to encode and to represent such motions. We investigate several applications and show results that illustrate the interest of such representation.

Index Terms—Simulation, Modeling, and Visualization; Computer Applications to Biology; Robotics.

1 Introduction

Nowadays, the dynamic nature of biological macromolecules, opposed to the static picture provided by X-ray crystallography, is generally accepted. Furthermore, it has been shown that flexibility plays key roles in molecular interactions such as protein-ligand [1], [2] and protein-protein docking [3], [4]. Unfortunately, and despite great advances achieved in the last years [5], [6], an atomic-resolution structural description of slow-timescale (large-amplitude) molecular motions is out of reach for currently available experimental methods [7]. Computational methods are therefore necessary to complement experimentation.

Molecular dynamics (MD) [8], [9] is the most widely used computational method to simulate molecular motions. MD is an appropriate method to analyze motions taking place in a short timescale (up to some nanoseconds). However, it is too computationally expensive for routine simulations of large-amplitude motions of macromolecules, even using coarse-grained models [10]. In some cases, MD simulations can be accelerated by the introduction of artificial forces [11]. Nevertheless, devising such forces may require prior knowledge of the particular problem, and they can excessively bias the resulting trajectories. Simulation methods based on Monte Carlo (MC) algorithms [8], [9] have been developed to overcome the limitations of MD. Such methods present however a major drawback for computing large-amplitude motions since the conformational exploration tends to get trapped into the many local minima of the complex molecular energy landscape. Alternatives to MD and MC simulations have been proposed using very different methods such as iterative NMA calculations [12], [13], or structurally constrained conformational exploration using models from rigidity theory [14].

Our method is based on path planning algorithms [15], [16], originally developed in the field of robotics. Such algorithms are efficient tools for exploring constrained high-dimensional spaces. Applied to problems in structural biology, they yield high-performance conformational search methods, able to consider a wide range of degrees of freedom. Like MC algorithms, they use random sampling to face the curse of dimensionality. However, they hold better coverage properties, and show less tendency to get trapped in local minima. In the recent years, path-planning-based methods have been successfully applied for investigating different problems such as: protein-ligand access and docking [17]–[19], protein and RNA folding [20]–[22], protein loop motions [23], domain motions [24], and motions of pairs of α-helices in transmembrane proteins [25].

This paper[1] recalls our approach for computing molecular motions (Section 2.2) and introduces voxel maps as a new and general computational tool to encode these motions (Section 2.3). Section 3 illustrates the potential interest of the method on several structural biology problems. The presented results show how voxel maps can effectively represent relative motions of two molecules, as well as conformational changes in proteins. The simplest application, presented in Section 3.1, consists in using voxel maps to identify channels in proteins, by exploring and encoding possible motions of a single atom between the active site and the surface. The second application (Section 3.2) addresses protein-ligand interactions. Voxel maps permit to reflect differences between the access/exit pathways of different ligands to the active site of a protein. Finally, Section 3.3 deals with the representation of conformational changes involving loop and domain motions.

2 Methodology

2.1 Overview

The method presented in this paper builds on the two-stage approach proposed in [19] for computing large-amplitude molecular motions. The first and main stage

J. Cortés and T. Siméon are with the LAAS-CNRS, F-31077 Toulouse, France
{jcortes,nic}@laas.fr
S. Barbe is with the LISBP, F-31077 Toulouse, France
sbarbe@insa-toulouse.fr
M. Erard is with the IPBS, F-31077 Toulouse, France
monique.erard@ipbs.fr
All the authors are with the Université de Toulouse; UPS, INSA, INP, ISAE; F-31077 Toulouse, France

[1] A preliminary version of this work was presented at the conference ICRA'09 [26].

consists in a geometric processing of the strongest molecular constraints (no atom overlaps, no bond breaking). Fast geometric computation [27] combined with efficient path planning algorithms [28] permits our method to generate large-amplitude motions of flexible molecules with very low computational cost. Optionally, in a second stage, results of the geometric exploration can be refined and analyzed using classic molecular modeling tools (e.g. energy evaluation/minimization), or using a path clustering technique [29] to extract the most energetically favorable motions from representative paths of the highest-score clusters.

The voxel-map representation described below can be seen as an intermediate layer between the two stages. It permits to arrange the information obtained from the exploration of a high-dimensional space (the molecular conformational space) into a simple three-dimensional data structure. The choice of the three dimensions and the size of voxels depends on the application (see Section 3). In addition to the information structuring, voxel maps permit a visual analysis of the results of the conformational exploration.

2.2 Exploring geometrically feasible motions

The conformational search method applied in this work (described in more detail in [19]) is based on a mechanistic modeling of molecules [30]. Groups of bonded atoms form the bodies of the mechanism, which are linked by articulations corresponding to bond torsions. These torsions are the molecular degrees of freedom. The atoms are represented by rigid spheres with (a percentage of) van der Waals radii[2]. These spheres cannot overlap. Structural features of the molecules can be translated into kinematic constraints in the mechanistic model. For instance, kinematic loop-closure constraints are imposed to keep the extremities of a flexible protein loop fixed [23]. Additional distance and orientation constraints can be imposed between elements of this mechanistic model in order to simulate interactions such as hydrogen bonds.

The technique applied to explore feasible motions of the mechanistic molecular model is derived from the *Rapidly-exploring Random Trees* (RRT) algorithm [32]. The basic principle of RRT is to incrementally grow a random tree, rooted at a given initial conformation q_{init}, to explore the search space for finding feasible paths. The exploration process is illustrated in Figure 1-b, and Algorithm 1 gives the pseudo-code for one iteration of the RRT construction. At each iteration, the tree is expanded toward a randomly sampled conformation q_{rand}. This random sample is used to simultaneously determine the tree node to be expanded and the motion direction. Given a distance metric in the search space (e.g. RMSD), the nearest node q_{near} in the tree is selected. Then, q_{near} is expanded towards q_{rand} by following a *local path* computed from the linear interpolation between the two points while the motion satisfies all the geometric constraints. If the expansion is feasible (i.e. it is possible to move more than a given ϵ), it leads to the generation of a new node q_{new} and a feasible local path p_{new}. The key feature of the RRT expansion strategy is to bias the exploration toward unexplored regions before uniformly covering the space.

Algorithm 1: ExpandRRT

```
input   : the current tree $\tau$, the model $M$;
output  : a new node $q_{new}$, a new edge $p_{new}$;
begin
    $q_{rand} \leftarrow$ SampleConf($M$);
    $q_{near} \leftarrow$ BestNeighbor($\tau$, $q_{rand}$);
    $q_{new} \leftarrow$ Expand($q_{near}$, $q_{rand}$);
    if not TooSimilar($q_{near}$, $q_{new}$) then
        $p_{new} \leftarrow$ SetEdge($q_{near}$, $q_{new}$);
        AddNodeToTree($\tau$, $q_{new}$);
        AddEdgeToTree($\tau$, $p_{new}$);
end
```

The RRT algorithm can be extended to treat mobile systems involving kinematic loop-closure constraints as detailed in [33]. Such an extension requires specific sampling functions that manage loop closure. An efficient geometric algorithm for sampling protein loop conformations is described in [23].

The algorithmic variant used in this work is ML-RRT [28], which has been shown to perform better when handling flexible molecular models. This variant considers two sets of conformational parameters: *active* and *passive*. Active parameters are essential for the system motion, and they are directly treated at each iteration of the algorithm, as explained above. Passive parameters, however, only need to be treated when they hinder the motion of active parts (i.e. the expansion of active parameters). In the applications presented in Section 3, the passive parameters correspond to the torsion angles of the protein side-chains, while the active parameters correspond to the other variables: the location and the internal torsions of the ligand, and the torsion angles of flexible protein backbone segments. The main advantage of ML-RRT is a much higher efficiency for dealing with high-dimensional problems thanks to the decoupled treatment of parameter subsets, which favors the exploration of active parameters.

2.3 Putting search trees into voxel maps

Nodes and edges of the RRT search tree are embedded in a high-dimensional space (the conformational space of the molecular model). The idea is to arrange this information into a lower-dimensional data structure for facilitating further analysis. The voxel-map representation

[2] Considering a percentage of the van der Waals equilibrium distance ensures that only energetically infeasible conformations are rejected by the collision checker. The value of 80% is often used in techniques that geometrically check atom overlaps [31].

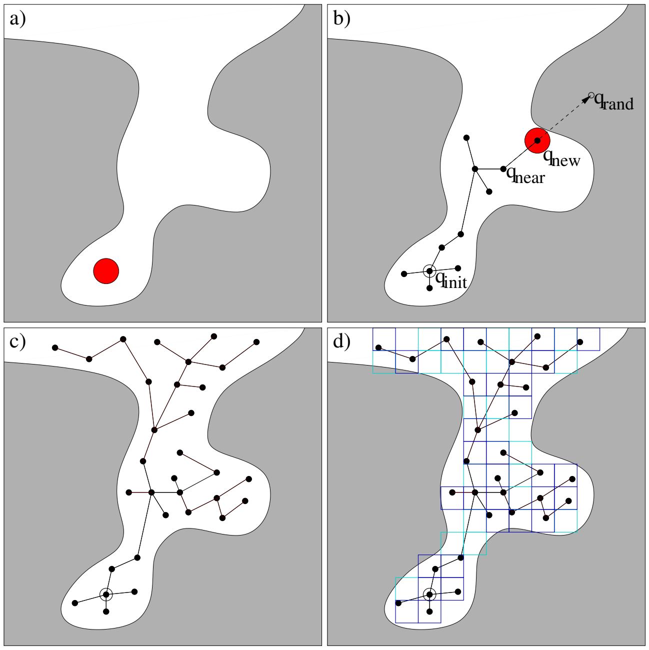

Fig. 1. Two-dimensional illustration of a voxel map construction. Given a geometric model of the molecules (a), a path planning algorithm is used to explore the subset of the conformational space that is reachable from the initial conformation satisfying motion constraints. b) Illustration of an expansion step of an RRT. c) Search tree resulting from the geometric exploration. d) Voxel map associated with the search tree.

Algorithm 2: ConstructVoxelMap

```
input  : the model $M$, the initial conformation $q_{init}$,
         the voxel map coordinates $v_{xyz}$, the voxel size
         $v_s$,
         the labeling function $f_{label}$;
output : the voxel map $\mathcal{V}$;
begin
    $\tau \leftarrow$ InitRRT($q_{init}$);
    $\mathcal{V} \leftarrow$ InitVoxelMap($q_{init}$, $v_{xyz}$, $v_s$);
    while not StopCondition($\tau, \mathcal{V}$) do
        $q_{new}$, $p_{new} \leftarrow$ ExpandRRT($\tau$, $M$);
        AddNodeToVoxelMap($\mathcal{V}$, $q_{new}$);
        AddEdgeToVoxelMap($\mathcal{V}$, $p_{new}$);
    LabelVoxels($\mathcal{V}$, $f_{label}$);
end
```

has been chosen for different reasons: (1) it is a simple and regular structure, which facilitates operations such as nearest-neighbor search, (2) it is three-dimensional, which permits a visual rendering of the information gathered during the conformational search.

The three dimensions of the voxel map correspond to three variables of interest for the particular problem. These three variables may be a subset of the conformational parameters of the mechanical model (e.g. position of the reference frame of a molecule, three selected bond torsions). They can also be chosen to encode information obtained from the conformation (e.g. position of one atom, center of mass of one domain). The voxel size also depends on the application, and on the chosen coordinates. Indeed, the resolution is chosen depending on the motion amplitude, and on the cost of the ulterior treatment (geometric and/or energetic analysis). Typically, in the applications presented below, voxel resolution (voxel edge length) varies from 0.1 Å to 2 Å. Note that increasing the voxel resolution does not affect computational efficiency. Most of the computational cost comes from the conformational exploration, whereas the generation of voxels from the resulting search tree is almost cost-less.

The process for generating the voxel map from the RRT search tree is very simple. Algorithm 2 gives the pseudo-code. The process is illustrated in Figure 1-d. At each expansion of the RRT search tree, the new node and the new edge are inserted into the voxel map. A new voxel is created if the projection of q_{new} on the voxel map coordinates lies in a yet uncovered region. Otherwise, the node is added to the list of nodes in the corresponding voxel. The procedure is slightly more complex for the edges. The new edge p_{new} is discretized and projected on the voxel map coordinate space. New voxels are created in yet uncovered traversed regions, and one intermediate conformation along the edge is associated to each of them. Intermediate edge conformations are also kept into existing traversed voxels, except for the voxels containing q_{new} and q_{near}. Thus, each voxel contains a list of conformations corresponding to the RRT nodes and edges that are projected on it.

The voxel map construction is iterated until a stop condition is satisfied. This stop condition can be based on a given size reached either in terms of the number of generated voxels or of the number of sampled conformations stored in the voxel map. Another possible variant is to stop if no new voxels are added after a given number of consecutive iterations, which reflects the difficulty to further explore new regions of the conformational space.

Once the voxel map has been constructed, voxels can be labeled in different ways, as illustrated by the applications described below. For instance, values can be assigned depending on the chronological order of generation. In this way, the voxel map can be used to display the regions of the space that are reached first during the conformational exploration (see Section 3.2). Other labeling procedures can be devised based on conformational or geometric features, such as the distance between catalytic residues (see the protein loop example in Section 3.3.1). The voxel labeling can also be made according to energy evaluation. An energetic analysis of the conformations associated with voxels may provide very useful information about the conformational energy landscape (see the example of protein domain motions in Section 3.3.2).

The method has been implemented within our software prototype BioMove3D. PyMOL [34] has been used for viewing molecular models and voxel maps. The computing times given in next section correspond to tests run on a single AMD Opteron 148 processor at 2.6 GHz.

3 Applications

The proposed approach - combining an RRT-based exploration method with the arrangement of the resulting conformations into a voxel map - was applied to examine dynamic properties of three relevant biological systems. The addressed problems involve the access/exit of ligands to buried active sites in proteins, and protein loop/domain motions. Note that in-depth explanations about the considered systems and a more detailed analysis of biological results are not the primary scope of this paper. The aim of this section is to show the potential interest of voxel maps in enhancing the understanding of biological problems involving molecular motions.

3.1 Findings channels in proteins

The most straightforward application of the method is the search and representation of channels in proteins. The channels are searched using the voxel-map technique described by Algorithm 2 to explore feasible motions of a ball (of arbitrary radius) inside the protein model. This algorithm permits to directly treat all side-chain flexibility[3] with a low computational cost. In this application, the three variables used for the voxel-map representation are the position parameters of the moving ball. The voxel resolution is chosen in relation to the ball size.

The benchmark for this application is cytochrome P450. The in/out channels of this enzyme have been recently characterized [35] using a computational technique called CAVER [36]. CAVER is based on the construction of a vertex-weighted graph from a discrete three-dimensional grid model of the protein. The weights are computed from the distance to the protein atoms, the lowest weights corresponding to nodes with the highest clearance. A variant of Dijkstra's algorithm is applied to search for the shortest low-cost paths. CAVER considers static structures, and performs systematic exploration of the protein interior. Molecular flexibility can only be indirectly treated by applying the technique to a set of structures (e.g. samples of a molecular dynamics simulation). The results described below aim to show some advantages of the voxel map method.

The structure represented in Figure 2 corresponds to bacterial P450-BM3 (PDBid 1JPZ). The voxel map in the figure represents the channels found by our method for a moving ball of radius 1.2 Å. Three of these channels, W, 2b and 2f, were also found by CAVER. However, channel 2d was not reported for this structure, although it was observed in a small number of other P450-BM3 structures [35]. A further analysis of channel 2d shows that the exit of the moving ball requires slight motions of some side-chains, in particular those of residues Leu20 and Leu29. This result shows the interest of considering side-chain flexibility when computing channels in proteins.

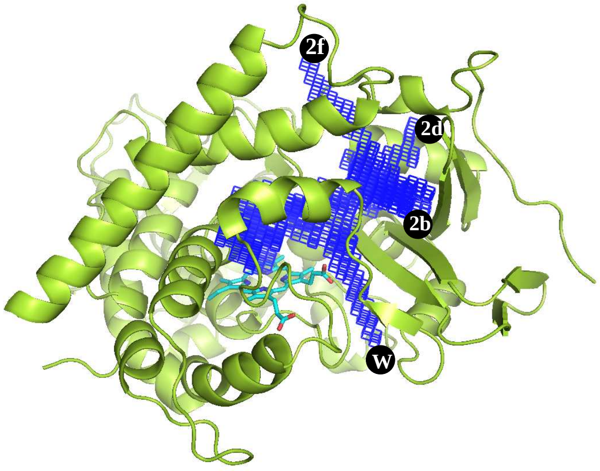

Fig. 2. Voxel-map representation of channels connecting the active site and the surface in bacterial P450-BM3. Channel identifiers follow the nomenclature in related work [35].

Also note that very similar channels are obtained from several runs of the voxel map computation with the moving ball initially located at different positions in the enzyme active site. Such a reliability shows the low sensibility to the starting point, which is another advantage over CAVER[4].

The computing time required to construct the voxel map representation of the channels was about 5 minutes. This is obviously not comparable to other available software such as CAVER [36], MOLE [37] or MolAxis [38], which only require some seconds to compute channels, but only treat rigid protein models. In spite of a lower computational efficiency, the incorporation of protein flexibility within our approach permits to identify previously closed channels in the protein model used as input. Besides, as illustrated below, our approach is able to consider flexible ligand models (instead of a single atoms) to compute access/exit channels. Furthermore, the voxel-map representation is not specially tuned for this specific application, and can be used to represent diverse molecular motions.

3.2 Analyzing ligand access/exit pathways

The proposed approach can be used to investigate the access/exit pathways of ligands (or substrates/products) to the active site of a protein. Such pathways can play

[3] The implementation of ML-RRT applied in this work only considers side-chain flexibility. Note however that we are currently developing an extension of ML-RRT that permits to treat flexible backbone regions such as protein loops.

[4] Results reported in [35] indicate that CAVER presents a significant sensibility to the starting point.

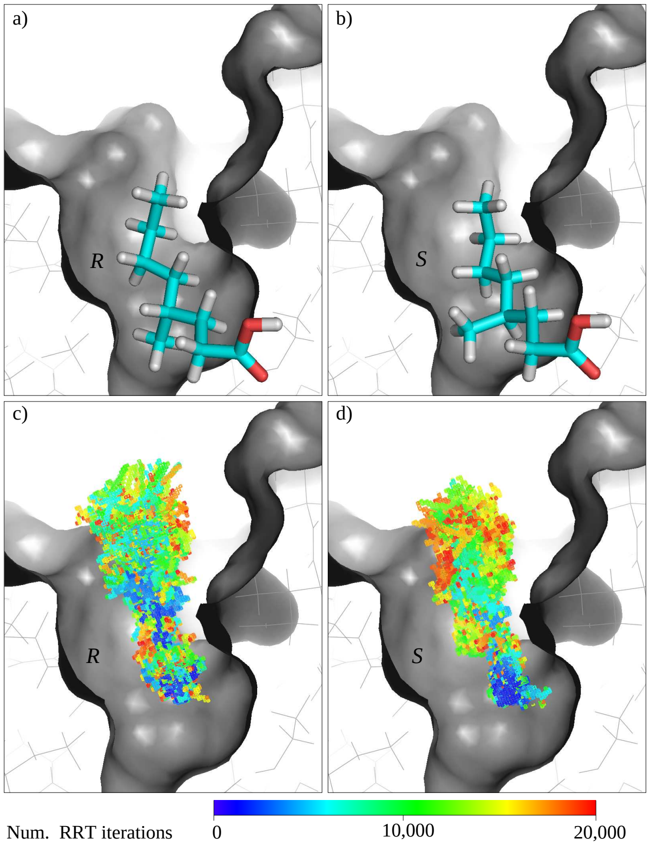

Fig. 3. a)-b) Models of the (*R*,*S*)-enantiomers of 4-methyloctanoic acid in the active site of CALB [40]. c)-d) Voxel maps representing locations of the center of mass of the (*R*,*S*)-enantiomers reachable from the catalytic position. Voxels resolution is 0.1 Å, and colors indicate the chronological order of generation.

key roles on the activity, specificity and selectivity of proteins presenting a deep and narrow binding pocket. For instance, the accessibility of substrates to the buried active site of an enzyme may influence its enantioselectivity [39].

To further analyze the relationship between the topology of the active site access channel and enzyme enantioselectivity, geometrically feasible motions of an enantiomer pair into the catalytic pocket of a lipase were explored using the voxel-map approach. The biological system chosen for this study is *Candida antarctica* lipase B (CALB), which is known to be enantioselective toward an enantiomer pair of (*R*,*S*)-4-methyloctanoic acid. This enzyme catalyzes preferentially the esterification of the (*R*)-enantiomer [40].

Geometrically feasible motions of the *R* and *S* substrates from its catalytic position [40] (Figure 3-a,b) were explored using the ML-RRT algorithm [28] within the voxel-map construction (Algorithm 2), which permits to directly treat the flexibility of the substrate and all protein side-chains with a low computational cost. The computed voxel maps represent the feasible positions reached by the substrate center of mass during the conformational exploration (Figure 3-c,d). The voxel resolution is 0.1 Å. Voxels have been colored depending on the chronological order of generation. Dark-blue voxels correspond to the positions that are reached first. These voxel maps reveal significant differences between behaviors of both enantiomers into the enzyme catalytic pocket. First, the root of the voxel map is notably narrower for the (*S*)-enantiomer (Figure 3-d) than for the (*R*)-enantiomer (Figure 3-c). This reflects the more constrained motions that the (*S*)-enantiomer must undergo to access and dock in a productive way at the catalytic site. Secondly, for the (*R*)-enantiomer, dark-blue voxels reach the middle-part of the map while such voxels are concentrated in the bottom part of the map for the (*S*)-enantiomer. The meaning is that, due to the spatial geometric constraints, the acess/exit of the (*R*)-enantiomer can be faster. These results tend to indicate that the topology of CALB active site is better suited for facilitating the reaction with (*R*)-4-mehyloctanoic acid, which correlates with experimental kinetic data showing a significant preference of CALB for this enantiomer [40]. As indicated in previous work [39], these results highly suggest that the accessibility of the substrate to the catalytic site and the difficulty encountered by the substrate in adopting a productive conformation at the reaction site may influence enzyme enantioselectivity.

Complementarily, it is possible to do an interpretation of the voxel maps in terms of entropy. The investigations reported in [41] suggest that the substrate accessible volume within the active site can be correlated to transition state entropy. Thus, since the voxel map represents the region explored by the substrate during its access/exit pathways to the active site, the volume of this region is an indicator of the entropic component of the activation free energy. In our tests with CALB, the volumes of the voxel maps computed for the (*R*)- and the (*S*)-enantiomer is about 25 Å^3 and 15 Å^3 respectively. Thus, the largest accessible volume for the (*R*)-enantiomer will tend to indicate that its interaction with CALB is entropically more favorable.

Concerning computational performance, the construction of the voxel map (including the ML-RRT exploration process) took less than 7 minutes for each enantiomer. Note that, aiming to get a very good coverage of the space into the active site cavity, the ML-RRT search tree expansion process yielded voxel maps with more than 25000 voxels for the (*R*)-enantiomer, and 15000 voxels for the more constrained motions of the (*S*)-enantiomer. This result shows that the method remains computationally fast, even for an exhaustive exploration.

3.3 Representing loop/domain motions

The RRT-based conformational search method can also be applied to compute large-amplitude internal molecular motions such as loop and domain motions [19], [23], [24]. Integrating the voxel-map representation in this approach provides a new tool for analyzing such conformational transitions.

3.3.1 Loop motions

The example illustrated in Figure 4 concerns the "WDP loop" in *Yersinia* protein tyrosine phosphatase (PTPase). The movement of the WDP loop plays a central role in the PTPase-mediated catalytic process [42], [43]. An open conformation of this loop permits the substrate access to the protein active site. Then, the WDP loop has to adopt a closed conformation that brings the catalytic residue Asp356 to a specific location for protein-substrate interaction. Starting from the open conformation of PTPase [42] (PDB ID: 1YPT), the voxel-map construction algorithm was applied to explore the mobility of the WDP loop (residues 352-361). In this case, the backbone torsion angles of the residues in the loop are the active parameters for the conformational exploration conducted by ML-RRT, while all the protein side-chain torsion angles are the passive ones. A voxel map obtained from the conformational exploration is represented in the right part of Figure 4. It displays the positions reached by the Cα atom of the middle loop residue Glu357. Voxels resolution is 0.5 Å, and colors have been assigned depending on the distance between the referred atom and the Cα of Val407, which is located on the bottom of the binding pocket. These atoms were also chosen in a related work [43] to measure the WDP loop gating during molecular dynamics simulations. The distance in the initial crystal structure is 17 Å. The minimum and maximum distances obtained through the conformational exploration are 11 Å and 22 Å respectively. The WDP loop reaches conformations very similar to the one in the closed structure [42] (PDB ID: 1YTN), with Cα RMSD below 1 Å. Besides, the voxel map shows that the loop can adopt more open conformations, as also suggested by molecular dynamic simulations [43]. Interestingly, the voxel map presents a marked pipe-like shape. Such a shape indicates that the WDP loop in PTPase is "mechanically" designed to perform opening-closure motions, while lateral-motions are not likely. This mechanical predisposition may explain the rapid opening-closure WDP loop motions reported in [43]. Finally note that constructing the voxel map required about 5000 iterations of Algorithm 2, with a computing time of 5 minutes.

3.3.2 Domain motions

Similarly, the proposed method can be applied for analyzing protein domain motions. The example presented here concerns the POU domain of N-Oct-3 transcription factor. This DNA binding domain recognizes numerous AT-rich DNA sequences. The structure of the molecule, represented in Figure 5, comprises two distinct, highly conserved sub-domains, termed POUs and POUh, connected by a flexible linker [44]. Due to its remarkable plasticity, the N-Oct-3 POU domain can adopt different conformations and corresponding homodimerization patterns, based on different relative positionings of POUs and POUh sub-domains, and depending on the type of DNA target [45]. The free N-Oct-3 POU domain is also in a different conformation. Previous studies [46] suggested the existence of a continuous running from free to "pre-bound" N-Oct-3 POU conformations, and that regulatory DNA regions likely select pre-existing conformers. Therefore, this domain represents a relevant model-system to study macromolecular flexibility.

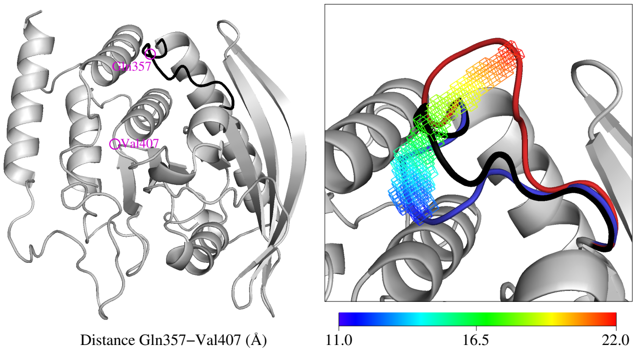

Fig. 4. (Right) Structure of PTPase (PDB ID: 1YPT), with the WDP loop in black color. (Left) Detail of the WDP loop and voxel map displaying the explored locations of the Cα atom of Glu357. Voxel colors indicate the distance between Cα atoms of Glu357 and Val407. The most open and closed loop conformations are represented in red and blue respectively.

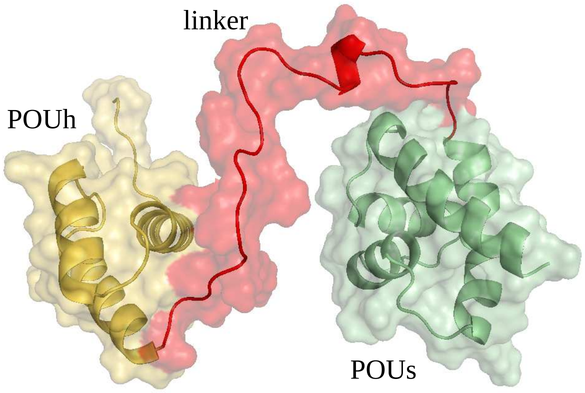

Fig. 5. Structure of N-Oct-3 POU domain.

The computational approach presented in this paper was applied in order to explore the conformational space of the N-Oct-3 POU domain, and to analyze the molecular mechanisms involved in the transitions between free and pre-bound conformations. The aim of the molecular modeling study was to explore possible locations of POUh sub-domain with respect to the POUs sub-domain. The conformational search was carried out by considering the flexibility of the long linker between both sub-domains (18 residues were considered to be fully flexible and 13 had limited flexibility), and all the protein side-chains being potentially flexible (i.e. their conformation changes if they hinder backbone

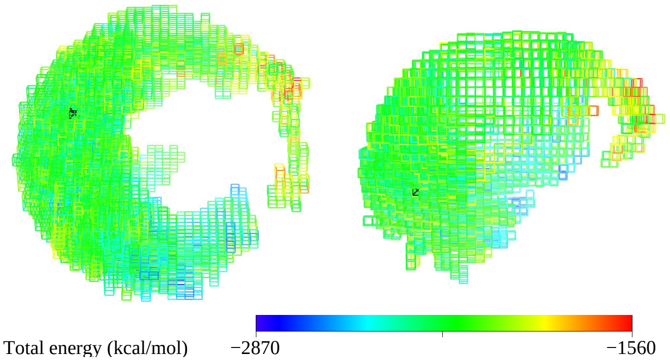

Fig. 6. Two views of the voxel map representing geometrically feasible motions of POUh with respect to POUs. The voxels display the relative position of the centers of mass of both domains. Voxels resolution is 2 Å, and colors have been assigned depending on the energies of the associated conformations. The black voxel indicates the initial conformation.

motions). Maximum and minimum values of the radius of gyration (data derived from SAXS experiments coupled to molecular modeling [46]) were integrated as constraints during the conformational exploration. The computed voxel map (Figure 6) represents the relative positions of the centers of mass of POUh with respect to POUs, thereby displaying possible locations of both subdomains. This voxel map construction required about 25000 iterations of Algorithm 2, with an overall computing time of 15 minutes. The 25000 sampled conformations were arranged in 5673 voxels of 8 Å^3 (voxel edge length = 2 Å). Conformations associated with each voxel were clustered into significantly different sets. Then, one conformation of each cluster was energy minimized[5] imposing constraints on the backbone atom positions in order to remain within the voxel. The voxel color was assigned according to the resulting lowest-energy conformation. Contrarily to the previous examples in which the computational cost of voxel labeling was very low, this energy-based labeling procedure is quite expensive. Indeed, the whole process required about 40 hours. Note however that, such a process is very easily parallelizable, and that voxel labeling can be restricted to regions of interest in order to avoid unnecessary waste of computational resources.

The volume of the computed voxel map, which represents reachable positions of POUh with respect to POUs, surpasses 45000 Å^3. Such a wideness of the explored conformational space reflects the high flexibility of the linker joining the sub-domains. This agrees with the critical importance ascribed to the linker with regard to the moulding of specific regulatory POU conformations to the target DNA [48], [49]. In addition, the voxel map enables the rapid identification of low-energy conformations attainable from the initial structure, and possible transition pathways between them. Bearing in mind the efficiency of the method, we plan to investigate a variety of domain descriptors and knowledge-based filters to optimize the biological relevance of the search.

4 Conclusions

We have presented a general framework for computing and representing molecular motions. The basic principle of the approach is to apply path-planning algorithms, originating from robotics, to explore feasible motions of mechanistic molecular models. The efficiency of such conformational exploration permits to attain large-amplitude motions with low computational cost. The voxel-map representation facilitates ulterior treatment of the explored conformations and permits direct visual interpretation of results. Besides, voxel maps could be used to bias or to focus the exploration to specific regions of the conformational space, and could permit to device more accurate metrics, considering motion feasibility, for improved conformational search algorithms.

In the presented work, 3D voxel maps have been mainly chosen for permitting visual display and analysis. However, the approach would be generalized to any dimension. In some cases, a more significant arrangement of structural information generated during the search could be done by choosing an arbitrary number of dimensions (and the associated parameters), which will be provided by dimensionality reduction methods such as PCA [50]. We consider this possibility for future work.

First results highlight the potential of the approach. Voxel maps can represent relative motions of two molecules. Such a representation displays the geometric suitability of a protein presenting a narrow, deep binding site for interacting with different ligands. It could be used to develop a predictive tool of enzyme enantioselectivity that would help to select a catalyst for a given racemate resolution. When applied to explore molecular deformations, voxel maps can provide a global representation of the conformational space of protein loops and protein domains undergoing large-amplitude motions. Such a representation would be very useful for the analysis and the prediction of macromolecular docking.

In conclusion, voxel maps can be seen as a general tool that, combined with other computational and experimental methods, will help to investigate the importance of flexibility and motion in molecular interactions.

Acknowledgments

The authors thank Maurice Franssen, David Guieysse and Magali Remaud for providing models and information about CALB. This work has been partially supported by the Région Midi-Pyrénées under projets ITAV ALMA and CTP AMOBIO, and by the French National Agency for Research (ANR) under project NanoBioMod.

[5] AMBER ff03 force field [47] has been used for the energetic analysis.

REFERENCES

[1] H. Carlson, “Protein flexibility is an important component of structure-based drug discovery,” *Curr. Pharm. Des.*, vol. 8, pp. 1571–1578, 2002.

[2] C. Cavasotto and N. Singh, “Docking and high throughput docking: Successes and the challenge of protein flexibility,” *Curr. Comput. Aided. Drug. Des.*, vol. 4, pp. 221–234, 2008.

[3] J. Janin, “Assessing predictions of protein-protein interaction: the CAPRI experiment,” *Protein Sci.*, vol. 14, pp. 278–283, 2005.

[4] L. Ehrlich, M. Nilges, and R. Wade, “The impact of protein flexibility on protein-protein docking,” *Proteins*, vol. 58, pp. 126–133, 2005.

[5] G. Katona, P. Carpentier, V. N. , P. Amara, V. Adam, J. Ohana, N. Tsanov, and D. Bourgeois, “Raman-assisted crystallography reveals end-on peroxide intermediates in a nonheme iron enzyme,” *Science*, vol. 316, pp. 449–453, 2007.

[6] P. Schanda, V. Forge, and B. Brutscher, “Protein folding and unfolding studied at atomic resolution by fast two-dimensional nmr spectroscopy,” *PNAS*, vol. 104, pp. 11 257–11 262, 2007.

[7] K. Henzler-Wildman and D. Kern, “Dynamic personalities of proteins,” *Nature*, vol. 450, pp. 964–972, 2007.

[8] A. Leach, *Molecular Modeling: Principles and Applications*. Cambridge: Longman, 1996.

[9] T. Schlick, *Molecular Modeling and Simulation - An Interdisciplinary Guide*. New York: Springer, 2002.

[10] E. Paci, M. Vendruscolo, and M. Karplus, “Validity of Gō models: Comparison with a solvent-shielded empirical energy decomposition,” *Biophys. J.*, vol. 83, pp. 3032–3038, 2002.

[11] S. Izrailev, S. Stepaniants, B. Isralewitz, D. Kosztin, H. Lu, F. Molnar, W. Wriggers, and K. Schulten, “Steered molecular dynamics,” in *Computational Molecular Dynamics: Challenges, Methods, Ideas. Vol. 4 of Lecture Notes in Computational Science and Engineering*, P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R. Skeel, Eds. Berlin: Springer-Verlag, 1998, pp. 39–65.

[12] L. Mouawad and D. Perahia, “Motions in hemoglobin studied by normal mode analysis and energy minimization: evidence for the existence of tertiary T-like, quaternary R-like intermediate structures,” *J. Mol. Biol.*, vol. 258, pp. 393–410, 1996.

[13] J. Jeong, E. Lattman, and G. Chirikjian, “A method for finding candidate conformations for molecular replacement using relative rotation between domains of a known structure,” *Acta. Cryst.*, vol. D62, pp. 398–409, 2006.

[14] S. Wells, S. Menor, B. Hespenheide, and M. Thorpe, “Constrained geometric simulation of diffusive motion in proteins,” *Phys. Biol.*, vol. 2, pp. 127–136, 2005.

[15] H. Choset, K. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. Kavraki, and S. Thrun, *Principles of Robot Motion: Theory, Algorithms, and Implementations*. Cambridge: MIT Press, 2005.

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

[17] A. Singh, J.-C. Latombe, and D. Brutlag, “A motion planning approach to flexible ligand binding,” *Proc. Conf. Intell. Syst. Mol. Biol. (ISMB)*, pp. 252–261, 1999.

[18] M. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, and J.-C. Latombe, “Stochastic conformational roadmaps for computing ensemble properties of molecular motion,” in *Algorithmic Foundations of Robotics V*, J.-D. Boissonnat, J. Burdick, K. Goldberg, and S. Hutchinson, Eds. Berlin: Springer-Verlag, 2004, pp. 131–147.

[19] J. Cortés, T. Siméon, V. Ruiz, D. Guieysse, M. Remaud, and V. Tran, “A path planning approach for computing large-amplitude motions of flexible molecules,” *Bioinformatics*, vol. 21, pp. i116–i125, 2005.

[20] N. M. Amato, K. A. Dill, and G. Song, “Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures,” *J. Comput. Biol.*, vol. 10, pp. 149–168, 2003.

[21] M. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, J.-C. Latombe, and C. Varma, “Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion,” *J. Comput. Biol.*, vol. 10, pp. 257–281, 2003.

[22] X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. Amato, “Using motion planning to study RNA folding kinetics,” *J. Comput. Biol.*, vol. 12, pp. 862–881, 2005.

[23] J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran, “Geometric algorithms for the conformational analysis of long protein loops,” *J. Comput. Chem.*, vol. 25(7), pp. 956–967, 2004.

[24] S. Kirillova, J. Cortés, A. Stefaniu, and T. Siméon, “An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins,” *Proteins*, vol. 70, pp. 131–143, 2008.

[25] A. Enosh, S. Fleishman, N. Ben-Tal, and D. Halperin, “Prediction and simulation of motion in pairs of transmembrane α-helices,” *Bioinformatics*, vol. 23, pp. e212–e218, 2007.

[26] J. Cortés, S. Barbe, M. Erard, and T. Siméon, “Encoding molecular motions in voxel maps,” *Proc. IEEE Int. Conf. Robotics and Automation*, 2009, in press.

[27] V. Ruiz de Angulo, J. Cortés, and T. Siméon, “BioCD: An efficient algorithm for self-collision and distance computation between highly articulated molecular models,” in *Robotics: Science and Systems*, S. T. snd G. Sukhatme, S. Schaal, and O. Brock, Eds. Cambridge: MIT Press, 2005, pp. 6–11.

[28] J. Cortés, L. Jaillet, and T. Siméon, “Disassembly path planning for complex articulated objects,” *IEEE Transactions on Robotics*, vol. 24, pp. 475–481, 2008.

[29] A. Enosh, B. Raveh, O. Furman-Schueler, D. Halperin, and N. Ben-Tal, “Generation, comparison, and merging of pathways between protein conformations: Gating in K-channels,” *Biophys. J.*, vol. 95, pp. 3850–3860, 2008.

[30] M. Zhang and L. Kavraki, “A new method for fast and accurate derivation of molecular conformations,” *J. Chem. Inf. Comput. Sci.*, vol. 42(1), pp. 64–70, 2002.

[31] M. DePristo, P. de Bakker, S. Lovell, and T. Blundell, “Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles,” *Proteins*, vol. 51, pp. 41–55, 2003.

[32] S. M. LaValle and J. J. Kuffner, “Rapidly-exploring random trees: Progress and prospects,” in *Algorithmic and Computational Robotics: New Directions (WAFR2000)*, B. Donald, K. Lynch, and D. Rus, Eds. Boston: A.K. Peters, 2001, pp. 293–308.

[33] J. Cortés and T. Siméon, “Sampling-based motion planning under kinematic loop-closure constraints,” in *Algorithmic Foundations of Robotics VI*, M. Erdmann, D. Hsu, M. Overmars, and F. van der Stappen, Eds. Berlin: Springer-Verlag, 2005, pp. 75–90.

[34] W. DeLano, “The PyMOL molecular graphics system,” 2002, http://www.pymol.org.

[35] V. Cojocaru, P. Winn, and R. Wade, “The ins and outs of cytochrome P450s,” *Biochim. Biophys. Acta*, vol. 1770, pp. 390–401, 2007.

[36] M. Petřek, M. Otyepka, P. Banás, P. Košinová, J. Koča, and J. Damborský, “CAVER: A new tool to explore routes from protein clefts, pockets and cavities,” *BMC Bioinfo.*, vol. 7, pp. 316–324, 2006.

[37] M. Petřek, P. Košinová, J. Koča, and M. Otyepka, “MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels,” *Structure*, vol. 15(11), pp. 1357–1363, 2007.

[38] E. Yaffe, D. Fishelovitch, H. Wolfson, D. Halperin, and R. Nussinov, “MolAxis: Efficient and accurate identification of channels in macromolecules,” *Proteins*, vol. 73, pp. 72–86, 2008.

[39] D. Guieysse, J. Cortés, S. Puech-Guenot, S. Barbe, V. Lafaquière, P. Monsan, T. Siméon, I. André, and M. Remaud-Siméon, “A structure-controlled lipase enantioselectivity investigated by a path planning approach,” *ChemBioChem*, vol. 9, pp. 1308–1317, 2008.

[40] M. Franssen, M. Kamp, L. Alessandrini, M. Huibers, and J. Vervoort, “The interaction between Candida antarctica lipase and branched chain fatty acids: a kinetic and molecular modelling study,” *Proc. Int. Symp. Biocatalysis and Biotransformations (Biotrans)*, 2005.

[41] J. Ottosson, L. Fransson, and K. Hult, “Substrate entropy in enzyme enantioselectivity: An experimental and molecular modeling study of a lipase,” *Protein Sci.*, vol. 11, pp. 1462–1471, 2002.

[42] J. Stuckey, H. Schubert, E. Fauman, Z.-Y. Zhang, J. E. Dixon, and M. Saper, “Crystal structure of Yersinia protein tyrosine phosphatase at 2.5 Å and the complex with tungstate,” *Nature*, vol. 370, pp. 571–575, 1994.

[43] X. Hu and C. Stebbins, “Dynamics of the WPD loop of the Yersinia protein tyrosine phosphatase,” *Biophys. J.*, vol. 91, pp. 948–956, 2006.

[44] D. Latchman, “POU family transcription factors in the nervous system,” *J. Cell. Physiol.*, vol. 179, pp. 126–133, 1999.

[45] R. Alazard, M. Blaud, S. Elbaz, C. Vossen, G. Icre, G. Joseph, L. Nieto, and M. Erard, “Identification of the ’NORE’ (N-Oct-

3 responsive element), a novel structural motif and composite element," *Nucleic Acids Res.*, vol. 33, pp. 1513–1523, 2005.

[46] R. Alazard, L. Mourey, C. Ebel, P. Konarev, M. Petoukhov, D. Svergun, and M. Erard, "Fine-tuning of intrinsic N-Oct-3 POU domain allostery by regulatory DNA targets," *Nucleic Acids Res.*, vol. 35, pp. 4420–4432, 2007.

[47] Y. Duan, C. Wu, S. Chowdhury, M. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, "A point-charge force field for molecular mechanics simulations of proteins," *J. Comput. Chem.*, vol. 24, pp. 1999–2012, 2003.

[48] W. Herr and M. A. Cleary, "The POU domain: Versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain," *Genes Dev.*, vol. 9, pp. 1679–1693, 1995.

[49] H. C. Leeuwen, M. J.Strating, M. Rensen, W. de Laat, and P. C. van der Vliet, "Linker lengthand composition influence the flexibility of Oct-1 DNA binding," *Embo J.*, vol. 16, pp. 2043–2053, 1997.

[50] M. Teodoro, G. P. Jr., and L. Kavraki, "Understanding protein flexibility through dimensionality reduction," *J. Comput. Biol.*, vol. 10, pp. 617–634, 2003.