An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins

Svetlana Kirillova, Juan Cortés, Alin Stefaniu, Thierry Simeon

TITLE

An NMA-Guided Path Planning Approach for Computing Large-Amplitude Conformational Changes in Proteins

SHORT TITLE

NMA-Guided Path Planning Approach

KEYWORDS

Protein flexibility, large-amplitude conformational transitions,
elastic network normal mode analysis, path planning algorithms,
adenylate kinase open-closed transition pathway.

AUTHORS

Svetlana Kirillova, Juan Cortés, Alin Stefaniu, Thierry Siméon

AFFILIATION (for all the authors)

LAAS-CNRS, Toulouse, France

CORRESPONDING AUTHOR

Juan Cortés
LAAS-CNRS, 7 Av. du Colonel Roche, 31077 Toulouse, France
Phone: +33 5 61 33 63 45
Fax: +33 5 61 33 64 55
E-mail: juan.cortes@laas.fr

ABSTRACT

This paper presents a new method for computing macromolecular motions based on the combination of path planning algorithms, originating from robotics research, and elastic network normal mode analysis. The low-frequency normal modes are regarded as the collective degrees of freedom of the molecule. Geometric path planning algorithms are used to explore these collective degrees of freedom in order to find possible large-amplitude conformational changes. To overcome the limits of the harmonic approximation, which is valid in the vicinity of the minimum energy structure, and to get larger conformational changes, normal mode calculations are iterated during the exploration. Initial results show the efficiency of our method, which requires a small number of normal mode calculations to compute large-amplitude conformational transitions in proteins. A detailed analysis is presented for the computed transition between the open and closed structures of adenylate kinase. This transition, important for its biological function, involves large-amplitude domain motions. The obtained motion correlates well with results presented in related works.

INTRODUCTION

The study of conformational changes in macromolecules such as proteins or DNA is of key importance for understanding their biological functions. Furthermore, accurately predicting molecular interactions with computational methods requires taking into account molecular flexibility[1–5]. Therefore, the development of techniques to compute macromolecular motions is currently the subject of much research.

The most rigorous method to simulate molecular motions is molecular dynamics (MD)[6,7], which calculates the trajectories of atoms using Newton's Second Law and potential energy functions of atom-atom interactions. However, the computational cost of MD prohibits routine simulations of large-amplitude motions of macromolecules. As an alternative to MD methods, stochastic search methods compute low-energy paths by randomly exploring a given molecular force field. Most of the stochastic approaches to compute molecular motions are based on Monte Carlo (MC) algorithms[6,7]. A recently proposed stochastic exploration method based on robotic path planning techniques[8] outperforms classical MC-based algorithms by simultaneously examining multiple pathways. Despite efforts to develop efficient techniques, the ability of energy-based search methods to compute large-amplitude motions of macromolecular models is limited by the complexity of the search-space. Indeed, the molecular energy landscape is a very high-dimensional manifold with many local minima.

The complexity of molecular energy landscapes led us to develop a two-stage approach for computing large-amplitude motions[9]. The first and main stage is purely geometric. A geometric treatment of the strongest molecular constraints, combined with efficient path planning algorithms, permits our method to consider large-amplitude motions with low computational cost. In the second stage, paths computed using the geometric approach are refined by fast energy minimization. We obtained encouraging results when applying this method to compute protein loop motions[10] and ligand-protein access pathways[9]. Recently, a similar approach has been used to compute motions of pairs of α-helices in transmembrane proteins[11]. However, despite the efficiency of such geometry-based conformational exploration, it remains difficult to directly handle fully-flexible molecular models with this approach because of the very high dimension of the conformational space.

This paper presents a new method to compute global macromolecular motions such as open-closed conformational transitions in proteins. It combines the above mentioned geometric approach with normal mode analysis (NMA)[12]. A number of works[13–16] have shown that large-amplitude motions in macromolecules (e.g. domain motions) are associated with low-frequency normal modes, therefore demonstrating the ability of NMA-based methods to predict the direction of collective conformational changes. However, such methods do not precisely represent the conformational transition. The harmonic approximation of the potential energy function is an accurate simplification only in the vicinity of the starting energy minimum. Computing large-amplitude transitions, which are complex inharmonic motions, requires additional tools. Previous works[17–20] have used an iterative process that recomputes the normal modes after each small displacement. The drawback of such a method is that the number of iterations necessary to compute a large-amplitude transition

can be very high. Since each normal mode calculation is computationally expensive, the process can be very time consuming. Thanks to the combination with the geometric path planning exploration, the approach presented in this paper requires only a small number of normal mode calculations. The main idea is to guide the conformational exploration, performed with a path planning algorithm on a purely geometric molecular model, using the directions of collective atomic motions given by the low-frequency normal modes. Indeed, the algorithm explores the space of the collective degrees of freedom provided by the low-frequency modes, which is a lower-dimensional sub-manifold of the conformational space.

MATERIAL AND METHODS

We first briefly describe the two basic methods used within our approach: the NMA method, which is based on an elastic network model of the molecule, and the geometric path search method. We then explain the algorithm for computing large-amplitude conformational changes that combines both methods.

Elastic Network Normal Mode Analysis

NMA is based on the harmonic approximation of the potential energy function around a minimum energy conformation. The diagonalization of the Hessian matrix of the potential energy provides a set of eigenvectors and eigenfrequencies that characterize the normal modes of vibration. The normal modes represent the atom displacements around the equilibrium position at the energy minimum. Collective motions, in which many atoms participate, are associated with low-frequency normal modes, while high-frequency modes correspond to local fluctuations.

In this work, we use an NMA-based method that considers a simplified potential for normal mode calculations based on an elastic network model (ENM) of the molecule[21]. The potential energy function of the ENM has the following form:

$$E_P = \sum_{d_{ij}^0 < R_c} \frac{C}{2}(d_{ij} - d_{ij}^0)^2$$

where d_{ij} is the distance between atoms i and j, and d_{ij}^0 represents their distance in the initial conformation. C is the elastic constant, which for simplicity is assumed to be the same for all interacting pairs. The sum is restricted to atom pairs separated by less than a distance R_c, which is an arbitrary cut-off parameter. The ENM is often applied to a coarse-grained molecular model that only considers one point per residue (e.g. the position of the C_α atoms). Such a simplified NMA variant, which we refer to as elastic network NMA (EN-NMA), can potentially predict the direction of collective conformational changes of proteins[15] and even larger macromolecular assemblies such as viruses[22]. Methods based on EN-NMA have been used for very different applications such as the evaluation of thermal fluctuations in proteins[23], flexible docking studies[24], flexible fitting of high-resolution macromolecular structures into low-resolution maps obtained by electron microscopy[19], and finding candidate conformation of multidomain proteins for use in molecular replacement[20]. The ENM has also been used within other methods to compute conformational transitions. Kim et al.[25] proposed a method for computing transition pathways using a coarse-grained ENM and interpolating the distances between spatially neighboring residues given by the two end conformations. The plastic network model (PNM) proposed by Maragakis and Karplus[26] also relies on the ENM. The elastic energy functions of the different conformers are involved in a single equation that is used to compute the conformational change pathway.

The most time consuming operation in normal mode calculations is the diagonalization of the Hessian matrix. Our current implementation integrates ElNémo[27]. This NMA

package considers all-atom models of molecules and applies a building-block approximation, called the rotation-translation block (RTB)[28], to speed up the calculation. The RTB considers rigid-body motions of atom groups (usually, one group per residue) instead of considering the individual motions of all the atoms. This approximation considerably reduces the size of the matrix, thus making the diagonalization much more efficient in terms of computing time and memory. It has been shown that the RTB approximation has very little influence in low-frequency modes[28].

A significant advantage of EN-NMA compared to NMA methods using detailed potentials is that the initial structure does not need to be energy minimized[21]. Indeed, any initial conformation is a minimum energy conformation for the ENM potential energy function. This advantage is particularly important for techniques, like the one presented in this paper, that repeatedly compute normal modes within an iterative process.

Geometric Path-Planning-Based Conformational Exploration

Path planning is a classical problem in robotics[29]. It consists of computing feasible motions for a mechanical system in a workspace cluttered with obstacles. In recent years, path planning techniques have undergone considerable progress. Sampling-based algorithms[30] have been demonstrated to be efficient tools for exploring constrained high-dimensional spaces. Such algorithms have been successfully applied to challenging problems in diverse application domains, including computational biology[8,9,11,31]. We present below an overview of the method described by Cortés et al.[9], which is extended in the present work to handle fully-flexible molecular models using collective motions provided by EN-NMA.

Mechanistic molecular model

Within our approach, molecules are modeled as articulated mechanisms. Groups of rigidly bonded atoms form the bodies of the mechanism and the articulations between bodies correspond to bond torsions. These torsions are the molecular degrees of freedom. The atoms are represented by spheres with (a percentage of) van der Waals radii. Fig. 1 shows the van der Waals sphere representation of adenylate kinase and illustrates the articulated mechanical model on a detailed view of a residue. Using a geometric interpretation of the van der Waals repulsive force, the spheres associated with non-bonded atom pairs cannot overlap. Additional distance and orientation constraints can be imposed between elements of this mechanistic molecular model in order to simulate attractive interactions such as hydrogen bonds or aromatic interactions.

We use the tailored algorithm BioCD[32] for efficient collision detection and distance computations in such mechanistic molecular models. BioCD uses spatially-adapted hierarchical data structures that approximate the shape of the protein at successive levels of detail, allowing the number of interacting pairs tested for collision to be significantly reduced. The algorithm is well suited to a sampling-based path planning scheme in which many degrees of freedom are simultaneously and arbitrarily modified during the exploration of the conformational space.

Conformational exploration algorithm

The conformational search technique applied in this work is based on the *Rapidly-exploring Random Trees* (RRT) algorithm[33]. RRT-like path planners perform well in highly constrained spaces. Molecular motions are generally extremely constrained due to steric clash avoidance.

The basic principle of the RRT algorithm is to incrementally grow a random tree rooted at the initial conformation $\mathbf{q}_{init}$ to explore the reachable conformational space and find a feasible path connecting $\mathbf{q}_{init}$ to a goal conformation $\mathbf{q}_{goal}$. Algorithm 1 gives the pseudocode for the RRT construction and Fig. 2 illustrates the process. At each iteration, the tree is expanded toward a randomly sampled conformation $\mathbf{q}_{rand}$, generated by the function `SamplePoint`. This random sample is used to simultaneously determine the tree node to be expanded and the direction in which it is expanded. The nearest node $\mathbf{q}_{near}$ in the tree to the sample $\mathbf{q}_{rand}$ is selected by the function `BestNeighbor` (given a distance metric in the conformational space) and an attempt is made by the function `ExpandTree` to expand $\mathbf{q}_{near}$ in the direction of the straight path to $\mathbf{q}_{rand}$. This straight-line connection is usually called a *local path*. A new node $\mathbf{q}_{new}$ is generated in the feasible sub-path (i.e. the local path segment satisfying motion constraints) between $\mathbf{q}_{near}$ and $\mathbf{q}_{rand}$. The process iterates until $\mathbf{q}_{goal}$ can be connected to the tree or a stop condition determines that no solution exists. Note that, in the absence of specified goal, the same algorithm can be used to encode within the computed tree a representative subset of feasible paths and conformations reachable from $\mathbf{q}_{init}$. The key idea of the RRT 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 any other node). Therefore RRT expansion is biased toward large Voronoi regions enabling rapid exploration before uniformly covering the space.

NMA-Guided Path Planning Method

The geometric path search method described above performs well when the macromolecular flexibility is only treated partially (e.g. flexible side-chains or flexible loops connecting rigid secondary structure elements)[9]. However, considering fully-flexible macromolecular models is much more complex because of the number of degrees of freedom and constraints. To address this complexity, the idea developed in this paper is to guide the conformational exploration carried out by the geometric approach using information about global molecular motions provided by low-frequency normal modes of vibration. This section first presents a brief description of the method and then explains the algorithmic details.

Outline of the method

To deal with fully-flexible molecular models, we define the search-space not as the space of the degrees of freedom of the mechanistic molecular model (i.e. the bond torsions) but as

the space of the low-frequency normal modes provided by EN-NMA. Each normal mode is considered as a collective degree of freedom and its amplitude is a parameter of the search-space. Starting from a given conformation, the RRT algorithm is used to explore the sub-space of the collective degrees of freedom that satisfies geometric constraints of the mechanistic model (i.e. collision avoidance). Thus, the computed tree encodes the feasible regions of the conformational space reachable from the initial conformation, following motion directions provided by a set of low-frequency normal modes while satisfying the constraints imposed on the geometric molecular model. Since the information provided by NMA is only accurate in a relatively small region around the initial conformation, the guidance of the RRT search would degrade when exploring larger regions. Therefore, normal mode calculations are iteratively updated during the conformational exploration in order to compute large-amplitude conformational transitions.

Algorithm

Algorithm 2 summarizes the main stages of the iterative search process. Starting from the initial conformation, the normal modes are computed and the RRT algorithm (Algorithm 1) is applied to explore possible motions in the space of the collective degrees of freedom until the search tree cannot be sensibly expanded toward unexplored regions of the search-space or until a given cost function can not be significantly improved. The resulting tree is analyzed in order to select the new start conformation for the next iteration. Thanks to the use of EN-NMA, the intermediate start conformations do not need to be energy minimized before the normal mode recalculation. At the end of the process, the conformational transition pathway is obtained by the concatenation of the sub-paths computed throughout all the iterations. We detail below the main features of the NMA-guided RRT search method.

The search-space. The search-space is the coordinate space of the low-frequency normal modes, which we denote by $\mathbf{S}_{NM}$. We consider an arbitrary number n of the lowest-frequency modes (as in related works[19,20], we consider 10-30 modes) and we assume that all of them have the same frequency. Thus, any combination of these n normal modes is potentially explored. This is a reasonable assumption for the first few normal modes obtained with coarse-grained EN-NMA methods. The movement along each mode m_i is parameterized by a single real variable $a_i \in [-1, 1]$. Therefore, the search-space $\mathbf{S}_{NM}$ is the Cartesian product of n real intervals $[-1, 1]$, which is an n-dimensional smooth manifold. Each point in $\mathbf{S}_{NM}$ is defined by a vector $\boldsymbol{\alpha} = \{a_1, \ldots, a_n\}$.

Conversion into internal coordinates. Most EN-NMA techniques, as the one used in this work, generate relatively small atom displacements in Cartesian coordinates following the directions given by the normal modes. Since normal modes are simply used as a guide for the conformational exploration, an amplification factor can be applied to these displacements disregarding its physical meaning. However, large atom displacements in Cartesian coordinates yield unrealistic bond lengths and bond angles. Thus, the displacements are

converted from Cartesian coordinates into internal coordinates, and the amplification factor is applied only to the bond torsions. In this way, each point $\boldsymbol{\alpha}$ in $\mathbf{S}_{NM}$ corresponds to a conformation of the mechanistic molecular model, $\mathbf{q}= f(\boldsymbol{\alpha})$. The function `SamplePoint` in Algorithm 1 generates conformations by randomly sampling values for $\boldsymbol{\alpha}$. Each node in the search tree stores $\mathbf{q}$, the molecular conformation, and $\boldsymbol{\alpha}$, the associated vector of parameters in the space of the normal modes.

Decoupled backbone and side-chain motions. Side-chain motions are disregarded by EN-NMA methods considering coarse-grained molecular models. In the case of ElNémo, even if an all-atom model is considered for computing the potential energy, the building-block approximation used for the diagonalization of Hessian matrix only considers one block per residue, therefore neglecting side-chain fluctuations. However, side-chain motions can be important during a conformational transition and must be taken into account for accurately computing the pathway. Our exploration method considers decoupled backbone and side-chain motions. The backbone atoms move following the directions provided by the combination of normal modes, while the side-chains only move when they collide. The procedure to compute side-chain motions is further explained in the paragraph below, describing the method used to validate local paths. Note that the different natures of backbone and side-chain motions in proteins have been experimentally shown by Lindorff-Larsen et al.[34]. On average, a protein can be characterized as having solid-like rigidity in the backbone with liquid-like side-chains attached.

Motion validation. A key process in the RRT algorithm is the expansion of a node in the search tree $\{\mathbf{q}_{near},\boldsymbol{\alpha}_{near}\}$ toward a randomly sampled point $\mathbf{q}_{rand} = f(\boldsymbol{\alpha}_{rand})$, yielding the generation of a new node $\{\mathbf{q}_{new},\boldsymbol{\alpha}_{new}\}$. This process is carried out by the function `ExpandTree` in Algorithm 1. A local path is generated from the linear interpolation between $\boldsymbol{\alpha}_{near}$ and $\boldsymbol{\alpha}_{rand}$. Beginning from $\mathbf{q}_{near}$, the satisfaction of motion constraints (i.e. collision avoidance) is verified for each conformation $\mathbf{q} = f(\boldsymbol{\alpha})$ along the local path. This path is validated using discrete steps. The discretization step size is determined using a conservative approach that guarantees that, at each step, no atom moves more than a prespecified distance. A collision between backbone atoms means that the local path cannot continue in the direction toward $\mathbf{q}_{rand}$. However, when a side-chain is involved in a collision, the method randomly perturbs its torsion angles aiming to find a collision-free conformation that will permit the expansion process to continue. The random perturbation is iterated a predefined number of times before determining that the backbone conformation is not valid because of side-chain collisions. The new node $\{\mathbf{q}_{new},\boldsymbol{\alpha}_{new}\}$ is the point obtained at the end of the valid portion of the local path.

Biased conformational search. Information about the conformational change can be used to bias the exploration performed by the RRT algorithm. We introduce this bias when selecting the node to be expanded, $\{\mathbf{q}_{near},\boldsymbol{\alpha}_{near}\}$. This node selection is performed by the function `BestNeighbor` in Algorithm 1. In the standard RRT algorithm, the selected node

is the nearest neighbor of the random sample $\{\mathbf{q}_{rand},\boldsymbol{\alpha}_{rand}\}$ for a given distance metric in the search-space. In our case, the metric is simply the Euclidean distance in the space of the normal modes $\mathbf{S}_{NM}$. In order to bias the selection, this distance can be weighted by a factor w that reflects the preference of the node for future expansion. If the goal conformation $\mathbf{q}_{goal}$ is known, the weight is computed from the RMS distance (RMSD) between the conformation $\mathbf{q}$ associated with the node and this goal conformation as:

$$w = \mathrm{RMSD}(\mathbf{q}_{goal}, \mathbf{q})/\mathrm{RMSD}(\mathbf{q}_{goal}, \mathbf{q}_{init})$$

Note that, in some cases, the goal conformation is unknown, but information is available about one or several pairs or residues that should be closer (or farther) after the conformational transition. The weight can then be computed from the distance between these residue pairs.

Termination conditions. Each RRT exploration iterates until a stop condition is satisfied. The procedure implemented for the `StopCondition` in Algorithm 1 detects when the search tree cannot be sensibly expanded toward unexplored regions of the search-space. The method consists of measuring the volume of a simple convex hull (e.g. a bounding box) of the tree and stopping when this volume does not increase after a given number of iterations. However, if information about the goal conformation is available, a simpler condition can be used that stops the exploration when the above defined weight associated with the conformations is not improved. Once the RRT construction stops, the function `SelectNewStart` (in Algorithm 2) analyzes the resulting tree in order to select the start conformation for the next iteration. If the goal conformation is known, the conformation with smallest RMSD to it is selected. Otherwise, the new start conformation is chosen among the farthest nodes to the initial one. In the current implementation, only one conformation is selected. Nevertheless, the algorithm could be easily modified to re-iterate the exploration from several significantly different conformations. The main process stops when the function `EndConformation` in Algorithm 2 determines that either the new starting point is arbitrarily close to the given goal conformation or the pathway cannot be significantly extended.

Solution pathway. The result of one iteration of Algorithm 2 is a sequence of elementary motions, each obtained from a different linear combination of modes, which are combined with side-chain motions. The computed pathway is the concatenation of the sub-paths obtained after each iteration. This output is a continuous path that satisfies the geometric constraints of our all-atom molecular model (no bond stretching and bending, and no atom overlaps). The continuous solution pathway can be discretized into an arbitrary number of feasible intermediate conformations.

RESULTS

The proposed method has been applied to compute transition pathways involving large-amplitude domain motions between known conformations of several proteins. A detailed analysis is first presented for adenylate kinase and results are compared to those obtained with other computational methods. Then, results obtained with other proteins allow us to analyze the performance of the method when applied to different protein types.

Adenylate Kinase Conformational Transition

Adenylate kinase (ADK) is a good testbed for our method for several reasons: it is a well-known protein, the structures of its open and closed conformers have been experimentally determined, the transition between these conformers involves large-amplitude domain motions, and this conformational change has been studied with different computational methods[18,26]. In this section, we show results obtained with our method for the transition pathway from the open conformer of ADK (PDB code: 4AKE) toward the closed conformer (PDB code: 1AKE). The structure of ADK[35] is divided into three domains. Two of these domains, called the LID and the NMPbind domains in related literature, are involved in the conformational transition, while the main domain, called the CORE, remains unchanged[26,36]. The three domains of ADK are represented with different gray levels in Fig. 3: the clearest region at the right is the CORE, the darkest portion at the top is the NMPbind domain, and the left domain is the LID. Fig. 3.a and Fig. 3.c show the open and the closed conformers respectively. Fig. 3.b represents an intermediate conformation of the transition pathway.

Transition pathway

Computing the transition pathway of ADK with our approach required only 10 iterations of the NMA-guided RRT search algorithm (Algorithm 2). Fig. 4 shows the ribbon superposition of the open conformer and the 10 intermediate conformations obtained after each iteration toward the closed conformer. Remarkably, although the whole protein model is potentially flexible, only the two segments between residues 35-63 and 112-169 move significantly (more than 3 Å). These two segments nearly correspond with the NMPbind and the LID domains respectively. The analysis of the obtained pathway shows that the conformational transition occurs in two steps. During the first iterations, only the LID starts closing toward the CORE. The motion of the NMPbind domain becomes significant from the 5th iteration, when the RMSD from the open conformer approaches 7 Å. Fig. 3.b shows the conformation obtained after the 5th iteration. At this iteration, the LID is near its final conformation in the closed conformer, while the NMPbind domain is closer to its initial conformation than to its final one. The plot in Fig. 5 represents the displacement of the C_α atoms measured for each pair of consecutive conformations in the pathway. One can observe two darker regions in this plot: the biggest one, between iterations 1 and 4, corresponds with the LID motion, while the region around residue 50 and between iterations 5

and 7 corresponds with the pathway segment where the motion of the NMPbind domain is more significant. Note that such a two-step nature of the conformational transition of ADK has been described before by Maragakis and Karplus[26].

Fig. 6 shows the variation of the RMSD (computed from the C_α atom positions in the all-atom model) along the transition pathway. The distance between the open and the closed conformers is approximately 9 Å. The conformation obtained after the 10th iteration of our computational method is very similar to the closed conformer. The C_α RMSD between both conformations is 1.95 Å. Fig. 7 shows the superposition of both structures[a]. Note that, although the RMSD associated with the conformational transition of ADK is moderate because of the rigidity of the CORE domain, the LID and the NMPbind domains undergo large-amplitude motions. In the computed pathway, the C_α of Gly151 (in the LID domain) moves more than 25 Å and the C_α of Ala55 (in the NMPbind domain) moves more than 12 Å.

The variation of the strain energy along the computed transition pathway is displayed in Fig. 8. The strain energy was calculated for the coarse-grained model (only considering the C_α atoms) using an elastic constant $C = 2.0 \times 10^{-2}$ kcal/mol per Å^2 and a cutoff of 8 Å, as proposed by Maragakis and Karplus[26]. Like in the referred work, the strain energy with relation to the open conformer remains very low during most of the pathway. It only starts increasing rapidly from an RMSD of 7 Å.

Performance and parameter settings

Computing ADK conformational transition pathway required about 80 minutes of CPU time on an Intel Pentium 4 processor at 3.0 GHz. Fig. 9 shows the computing time spent in each iteration of the RRT-based conformational exploration. The plot also displays the number of nodes generated in each search tree. The normal mode calculation preceding each exploration step took about 1 minute. Note that, although the above results mainly concern the 10 intermediate conformations obtained after each iteration of Algorithm 2, the output of our computational method is a continuous smooth path, free of steric clashes for the all-atom protein model. A movie of this pathway is available as supplementary material.

We used the following parameter settings: we considered the 20 lowest-frequency normal modes provided by ElNémo for a cutoff distance of 8 Å; we applied an amplification factor of 100 to the bond torsions extracted from the atom displacements given by the normal modes; atom collisions were checked using spheres with 80% of the van der Waals radii[b]; the discretization step for validating the local paths permitted a maximum atom displacement of 0.1 Å; the RRT exploration stops when 100 new nodes are consecutively generated without reducing the RMSD.

[a] We have used MASS[37] for structure superposition.

[b] 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[38].

Other Examples

Besides ADK, the proposed method has been applied to several proteins selected from the Database of Macromolecular Movements (DMM)[39]. This section provides results obtained for four other proteins: ATP sulfurylase (ATP), C. Glutamicum DAP Dehydrogenase (DAP), EIAV Capsid Protein P26 (EIA), and Lysine/Arginine/Ornithine binding protein (LAO). All the examples involve large-amplitude domain motions: ATP, DAP, and LAO movements are classified predominantly hinge in the DMM, while EIA movement is considered as predominantly shear. Table I provides the PDB IDs corresponding to initial and final conformations and summarizes the numerical results obtained when computing the associated conformational transitions. All tests were performed using the same parameter settings as for ADK (see above). The values reported in the table for $\mathrm{RMSD}_{\mathrm{init}}$ and $\mathrm{RMSD}_{\mathrm{end}}$ are the RMSD to the goal conformation measured at the beginning and at the end of the computed pathways. In all the cases, the RMSD to the goal conformation is reduced below 2 Å. Also note that the number of iterations of the NMA-guided RRT search algorithm (Algorithm 2), $\mathrm{N}_{\mathrm{iter}}$, required to compute the transition pathway remains very small: it ranges from 3 for ATP to 12 for LAO. The total computing time[c], $\mathrm{T}_{\mathrm{total}}$, ranges from 70 minutes for EIA to 200 minutes for DAP. $\mathrm{T}^i_{\mathrm{NMA}}$ and $\mathrm{T}^i_{\mathrm{explor}}$ respectively represent the average computing time required for the normal mode calculation and the RRT-based conformational exploration at each iteration. Results reported in the table indicate that $\mathrm{T}^i_{\mathrm{NMA}}$ strongly depends on the protein size ($\mathrm{N}_{\mathrm{res}}$ is the number of residues in each protein), while $\mathrm{T}^i_{\mathrm{explor}}$ is much less affected by that. Indeed, $\mathrm{T}^i_{\mathrm{explor}}$ and $\mathrm{N}_{\mathrm{iter}}$ are more related to the protein structure and shape than to the protein size. Interestingly, in the two examples requiring a lower number of iterations (ATP and EAI) the domains are connected by a single linker, while two linkers exist between domains in the other cases (see Fig. 10 for illustration). When two domains are connected by two linkers, the N- and C-terminal regions are in the same domain. Working in internal coordinates, it is difficult to maintain the shape of the domain containing the terminal regions during the conformational exploration. Thus the feasible displacements generated by the RRT-based search are smaller in that case. For DAP, the interface between the two domains is a wider region than for LAO, and thus their relative motion is more constrained. This may explain the higher value of $\mathrm{T}^i_{\mathrm{explor}}$.

These results confirm the ability of the NMA-guided RRT search method to compute large-amplitude conformational changes of all-atom protein models from a very small number of NMA calculations. The number of required NMA calculations is notably lower when compared to existing iterative NMA-based methods[17–20], which are limited to computing small displacements of the protein structure at each iteration. The performance gain with respect to these other methods would still increase for large proteins, since the computing time associated with NMA calculations rapidly grows with the protein size.

[c]Computing time corresponds to runs on a single Intel Pentium 4 processor at 3.0 GHz.

DISCUSSION

On the direction of the conformational transition. Results presented above concern conformational transitions from open conformers toward closed conformers. Nevertheless, the method can also be applied to computing closed-to-open pathways, and our tests show that the search algorithm is not very sensitive to the chosen direction. Computing the transition pathway from the closed conformer toward the open conformer of ADK required 11 iterations of the NMA-guided RRT search algorithm (Algorithm 2) and about 80 minutes of computing time for attaining an RMSD of 1.97 Å. Fig. 11 shows the superposition of the open conformer and the conformation obtained after the 11th iteration. These results are very similar to the ones presented in the preceding section. The profile of the RMSD variation displayed in Fig. 12 is also very similar to the profile corresponding to the open-to-closed transition pathway. Our tests indicate that the low-frequency normal modes computed from the closed conformer are still a suitable guide for the RRT-based search method, even if EN-NMA methods have been shown to perform better when applied to an open protein form than to a closed form[15].

On the number of modes. We also analyzed the influence of the number of normal modes used as collective degrees of freedom within the NMA-guided RRT search on the performance of the method. Table II summarizes results obtained using the 5 and the 20 lowest-frequency modes for the first iteration of ADK transition pathway computation in both directions (open-to-closed and closed-to-open). The reported computing time T and the RMSD reduction have been averaged over 10 runs of the iteration. One can note that the number of modes has very little influence when starting from the open conformation. However, a higher number of modes is necessary to obtain a good performance of the algorithm when the exploration starts from a closed protein form. The different behavior is due to stronger motion constraints to avoid steric clashes for the closed form of the protein compared to the open one. In the latter case, since less constrained conformational space regions are explored, combinations of very few modes may be sufficient to yield feasible large-amplitude motions. In contrast, feasible motions are more complex near the closed form, and require a combination of a higher number of modes. Even for such constrained situations, tests (not reported in this paper) using a higher number of modes (up to 50) do not show significant performance improvement compared to tests using 20 modes. From our tests, using 20 modes presents the best compromise between exploration rate and computing time. Therefore, this value has been considered as a constant for all the iterations. During the search, all the lowest frequency modes are assumed to have the same frequency. Thus, any linear combination of them can be explored. This is a reasonable assumption since we only use the first few normal modes computed from a simplified ENM potential. However, if a higher number of modes was considered, it might be suitable to better guide the exploration by applying frequency-dependent weights to emphasize the lowest modes, as proposed by Jeong et al.[20].

On the method reliability. Despite the random nature of the RRT-based conformational exploration, different runs of the algorithm yield very similar transition pathways. The normal mode guidance combined with the geometric constraints drive the molecular motion toward a unique pathway class. Fig. 13 and Fig. 14 display results for 10 different runs of the first iteration of the algorithm starting from the open ADK conformer. The structure superposition in Fig. 13 shows that the obtained conformations are remarkably similar: the average C_α RMSD is 0.62 Å. The plot in Fig. 14 shows that each residue moves approximately the same distance in each run. This result demonstrates the reliability of the method.

On the convergence. Looking at Fig. 6 and Fig. 12, one can observe a slower convergence rate of the iterations when approaching the final conformation. The RMSD to the goal conformer decreases very rapidly during the first iterations, but later, it tends to show a slower asymptotic convergence. This is a common behavior for most iterative NMA methods[19]. The rigid geometry assumption for the internal coordinates considered in our approach may also explain the difficulty reducing RMSD to the goal conformation below 2 Å. Indeed, since the conformational exploration considers a protein model in which bond lengths and bond angles are maintained constant at the values of the initial structure, it is not possible to exactly converge to a goal conformation for which the values of these parameters are in general slightly different. Introducing variable bond lengths and bond angles in the articulated molecular model might enable the method to decrease the final RMSD, but will require more costly update operations that would slow-down the overall performance.

On side-chain motions. Side-chain motions may have important implications in the protein conformational change. Compared to other methods applying coarse-grained EN-NMA that consider rigid side-chains, a nice feature of the NMA-guided RRT search method is that it handles side-chain mobility. Indeed, the exploration modifies side-chain conformations during the transition pathway when they are involved in steric clashes. The efficiency of the current procedure, based on simple random perturbation of the side-chain torsions, although sufficient, may probably gain from using a rotamer library to bias the random sampling. Another possible improvement for better considering side-chain mobility may be to use all-atom normal mode calculations computed from detailed potentials instead of the EN-NMA method. Such a more accurate (but computationally expensive) method would better deal with coupled motions of the backbone and the side-chains. Note that the DIMB method[40] could be an interesting choice for such improvement because of its computational efficiency.

On predicting unknown candidate conformations. While the results above show the good performance of the approach for computing large-amplitude transition pathways between known conformers, the problem of predicting feasible pathways to unknown candidate conformations remains a much more challenging and yet open problem. Providing

an accurate predictive method is beyond the scope of this paper. However, the discussion below highlights the potential of the NMA-guided RRT search algorithm to address such problems. In the case of an unknown goal conformation, the RRT search tree tends to cover the subset of the conformational space reachable from the initial conformation. Starting from the open conformer of ADK, we applied this exploratory variant to the first iteration of the algorithm. The exploration stopped after 5 minutes with a tree containing almost 1000 nodes. The resulting tree nodes, ordered by increasing RMSD to the initial open conformation, were then clustered in a very simple way in order to select significantly different conformations. Conformations were only kept if the RMSD between them was greater than a given threshold. For a threshold equal to the maximum RMSD from the initial conformation to any node in the tree, only two significantly different conformations were obtained. Fig. 15 shows these two conformations superposed to the open ADK conformer. One can observe that in both cases the most mobile part is the LID domain, while the rest of the protein remains almost unchanged. This behavior corresponds to the beginning of the transition pathway presented in the preceding section. For the farthest conformation (Fig. 15 left), the LID domain moves toward its position in the closed conformer, and the RMSD to this conformation is reduced of 1.00 Å. Note that the other selected conformation (Fig. 15 right) corresponds to a LID motion in the opposite direction. This preliminary study tends to show that the approach could be extended for predicting pathways to unknown protein conformations. However, this interesting extension will require the introduction of some form of energetic scoring for selecting candidate conformations.

CONCLUSION

We have presented an efficient and reliable method for computing large-amplitude conformational changes in proteins. Combining normal mode analysis and geometric path planning algorithms, the approach overcomes the limitations of each individual method for computing this type of macromolecular motion: the normal modes are only able to accurately represent motions in a vicinity of the minimum energy structure, while geometric path planning approaches need guidance to obtain realistic motions of fully-flexible macromolecular models. The described procedure is completely automatic (i.e. does not need user intervention) and requires tuning of few parameters. The output of the proposed algorithm is a continuous pathway that satisfies the geometric constraints of a mechanistic all-atom molecular model. These constraints correspond to the strongest energetic barriers. The analysis of such a geometrically feasible pathway can provide very useful information. Our study of ADK conformational transition indicates that the solution obtained with this mechanistic approach correlates well with results of previous studies. Additionally, a rapid energy minimization of intermediate conformations may enable a finer analysis of the transition pathway. Our goal in the near future is to apply the method presented in this paper to consider protein flexibility in the computational analysis of protein-ligand interactions. We expect to extend our previous work on protein-ligand accessibility[9] to the case of fully-flexible receptor models.

ACKNOWLEDGMENTS

We would like to thank Liliane Mouawad for valuable discussions on normal mode analysis and for her help proofreading a previous version of this manuscript. This work has been partially supported by the LAAS-CNRS project AMoRo, the ITAV project ALMA, and the ANR project NanoBioMod.

REFERENCES

[1] Carlson HA. Protein flexibility is an important component of structure-based drug discovery. Curr Pharm Des 2002;8:1571-1578.

[2] Cavasotto CN, Orry AJW, Abagyan RA. The challenge of considering receptor flexibility in ligand docking and virtual screening. Curr Comput Aided Drug Des 2005;1:423-440.

[3] Janin J. Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci 2005;14:278-283.

[4] Ehrlich LP, Nilges M, Wade RC. The impact of protein flexibility on protein-protein docking. Proteins 2005;58:126-133.

[5] Floquet N, Marechal J-D, Badet-Denisot M-A, Robert CH, Dauchez M, Perahia D. Normal mode analysis as a prerequisite for drug design: application to matrix metalloproteinases inhibitors. FEBS Letters 2006;580:5130-5136.

[6] Leach AR. Molecular modeling: principles and applications. Essex: Longman; 1996.

[7] Schlick T. Molecular modeling and simulation - an interdisciplinary guide. New York: Springer; 2002.

[8] Apaydin MS, Brutlag DL, Guestrin C, Hsu D, Latombe J-C, Varma C. Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. J Comput Biol 2003;10:257-281.

[9] Cortés J, Siméon T, Ruiz-deAngulo V, Guieysse D, Remaud-Siméon M, Tran V. A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 2005;21:i116-i125.

[10] Cortés J, Siméon T, Remaud-Siméon M, Tran V. Geometric algorithms for the conformational analysis of long protein loops. J Comput Chem 2004;25:956-967.

[11] Enosh A, Fleishman SJ Ben-Tal N, Halperin D. Prediction and simulation of motion in pairs of transmembrane α-helices. Bioinformatics 2007;23;e212-e218.

[12] Cui Q, Bahar I. Normal mode analysis: theory and applications to biological and chemical systems. Boca Raton: CRC Press; 2006.

[13] Brooks BR, Karplus M. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. Proc Natl Acad Sci USA 1985;82:4995-4999.

[14] Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins 1998;33:417-429.

[15] Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode analysis. Protein Eng 2001;14:1-6.

[16] Alexandrov V, Lehnert U, Echols N, Milburn D, Engelman D, Gerstein M. Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. Protein Sci 2005;14:633-643.

[17] Mouawad L, Perahia D. 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 1996;258:393-410.

[18] Miyashita O, Onuchic JN, Wolynes PG. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc Natl Acad Sci USA 2003;100:12570-12575.

[19] Tama F, Miyashita O, Brooks III CL. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. J Struct Biol 2004;147:315-26.

[20] Jeong JI, Lattman EE, Chirikjian GS. A method for finding candidate conformations for molecular replacement using relative rotation between domains of a known structure. Acta Cryst 2006;D62;398-409.

[21] Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905-1908.

[22] Tama F, Brooks III CL. The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. J Mol Biol 2002;318:733-747.

[23] Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173-181.

[24] Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant normal modes. J Am Chem Soc 2005:127;9632-9640.

[25] Kim MK, Jernigan RL, Chirikjian GS. Efficient Generation of Feasible Pathways for Protein Conformational Transitions. Biophys J 2002;83:1620-1630.

[26] Maragakis P, Karplus M. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. J Mol Biol 2005;352: 807-822.

[27] Suhre K, Sanejouand YH. ElNémo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 2004;32:W610-W614.

[28] Tama F, Gadea FX, Marques O, Sanejouand YH. Building-block approach for determining low-frequency normal modes of macromolecules. Proteins 2000;41;1-7.

[29] Latombe J-C. Robot motion planning. Boston: Kluwer Academic Publishers; 1991.

[30] LaValle SM. Planning algorithms. New York: Cambridge University Press; 2006.

[31] Amato NM, Dill KA, Song G. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. J Comput Biol 2003;10:239-255.

[32] Ruiz de Angulo V, Cortés J, Siméon T. BioCD: an efficient algorithm for self-collision and distance computation between highly articulated molecular models. In: Thrun S, Sukhatme G, Schaal S, Brock O, editors. Robotics: Science and Systems I. Cambridge: MIT Press; 2005. p 6-11.

[33] LaValle SM, Kuffner JJ. Rapidly-exploring random trees: progress and prospects. In: Donald BR, Lynch KM, Rus D, editors. Algorithmic and computational robotics: new directions (WAFR2000). Boston: AK Peters; 2001. p 293-308.

[34] Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 2005;433:128-132.

[35] Müller CW, Schulz GE. Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 Åresolution. A model for a catalytic transition state. J Mol Biol 1992;224:159-177.

[36] Müller CW, Schlauderer GJ, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 1996;4:147-156.

[37] Dror O, Benyamini H, Nussinov R, Wolfson H. MASS: Multiple structural alignment by secondary structures. Bioinformatics 2003;19:i95-i104.

[38] DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41-55.

[39] Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ, Thorpe M, Gerstein M. The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res 2006;34:D296-301.

[40] Mouawad L, Perahia D. Diagonalization in a mixed basis: a method to compute low-frequency normal modes for large macromolecules. Biopolymers 1993;33:599-611.

Table I: Numerical results for ADK and four other proteins that undergo large-amplitude conformational transitions.

Protein	Init PDB ID	Goal PDB ID	N_{res}	$RMSD_{init}$	$RMSD_{end}$	N_{iter}	T^i_{NMA}	T^i_{explor}	T_{total}
ADK	4ake	1ake	214	8.73 Å	1.95 Å	10	75 s	400 s	80 min
ATP	1m8p	1i2d	572	7.01 Å	1.90 Å	3	1220 s	540 s	88 min
DAP	1dap	3dap	320	5.59 Å	1.89 Å	9	310 s	1025 s	200 min
EIA	1eia	2eia	206	5.17 Å	1.30 Å	6	70 s	625 s	70 min
LAO	2lao	1laf	238	8.19 Å	1.68 Å	12	100 s	780 s	176 min

Table II: Experiments on the influence of the number of modes over the performance of the computational method. Results for one iteration of the NMA-guided RRT search algorithm (Algorithm 2) in the two directions of ADK conformational transition.

ADK	**5 modes**		**20 modes**	
Transition	**T**	**ΔRMSD**	**T**	**ΔRMSD**
open → closed	390 s	1.22 Å	399 s	1.41 Å
closed → open	316 s	0.38 Å	395 s	1.71 Å

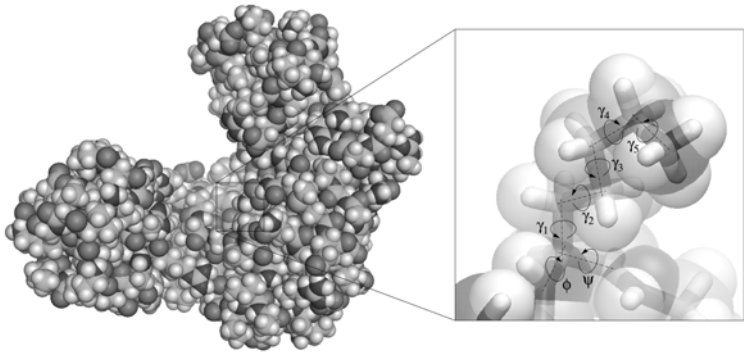

Figure 1: Mechanistic molecular model. Proteins are modeled as articulated mechanisms. Bonded atom groups form the bodies and the articulations correspond to bond torsions.

Algorithm 1: Construct_RRT

input : the search-space $\mathbf{S}$, the root $\mathbf{q}_{init}$
output: the tree τ
begin
  $\tau \leftarrow$ `InitTree`($\mathbf{q}_{init}$);
  while not `StopCondition`(τ) **do**
    $\mathbf{q}_{rand} \leftarrow$ `SamplePoint`($\mathbf{S}$);
    $\mathbf{q}_{near} \leftarrow$ `BestNeighbor`(τ, $\mathbf{q}_{rand}$);
    $\mathbf{q}_{new} \leftarrow$ `ExpandTree`($\mathbf{q}_{near}$, $\mathbf{q}_{rand}$);
    if not `Similar`($\mathbf{q}_{near}$, $\mathbf{q}_{new}$) **then**
      `AddNewNode`(τ, $\mathbf{q}_{new}$);
      `AddNewEdge`(τ, $\mathbf{q}_{near}$, $\mathbf{q}_{new}$);
end

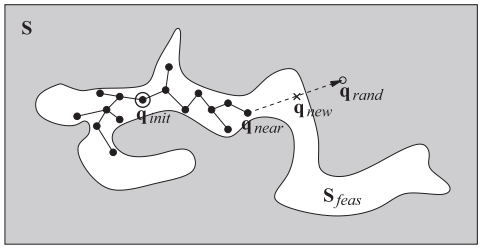

Figure 2: Illustration of one expansion step of a search tree using an RRT-based algorithm. The tree tends to cover $\mathbf{S}_{feas}$: the feasible subset of the search-space $\mathbf{S}$.

Algorithm 2: Compute_Pathway

```
input  : the initial conformation q_init
output: the transition pathway p
begin
    p ← InitPath(q_init);
    q_i ← q_init;
    while not EndConformation(q_i) do
        S_NM ← ComputeNormalModes(q_i);
        τ ← Construct_RRT(S_NM, q_i);
        q_n ← SelectNewStart(τ);
        p ← AddSubpath(τ,q_i,q_n);
        q_i ← q_n;
end
```

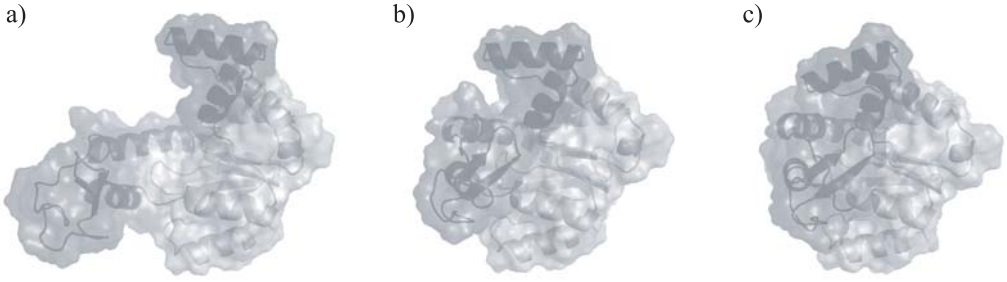

Figure 3: Three frames of the conformational transition of adenylate kinase: (a) open conformer, (c) closed conformer, (b) an intermediate conformation. The CORE (right), LID (left) and NMPbind (top) domains are represented with different gray levels.

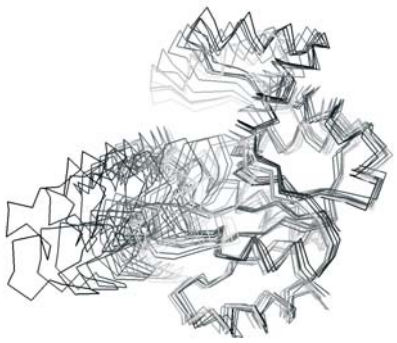

Figure 4: Superposition of the open ADK conformer (black) and 10 intermediate conformations (gray-scale) obtained after each iteration for computing the transition pathway toward the closed conformer.

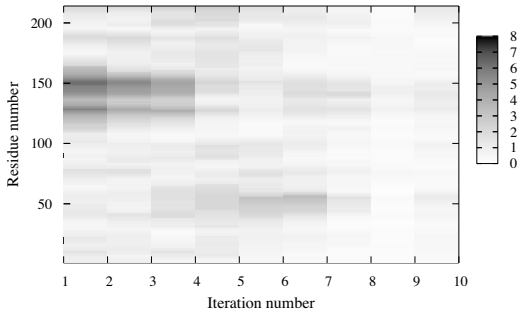

Figure 5: Displacement of the residues along the conformational transition. The plot shows the displacement of the C_α atoms for each iteration step along the pathway. The darker regions correspond with the more mobile segments.

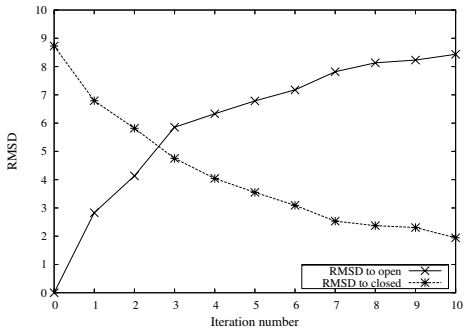

Figure 6: Variation of the RMSD along the computed pathway from the open conformer toward the closed conformer of ADK.

Figure 7: Superposition of the closed ADK conformer (black) and the final conformation of the computed transition pathway from the open conformer (gray).

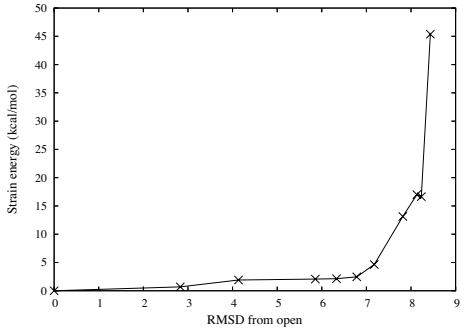

Figure 8: Variation of the strain energy along the conformational transition pathway.

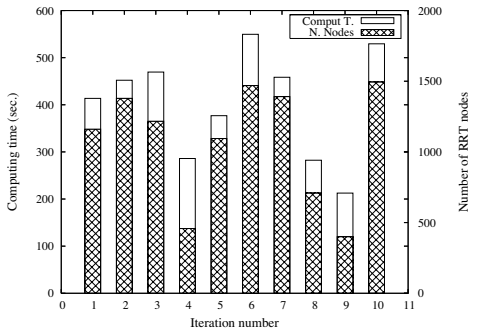

Figure 9: Performance of the conformational exploration technique: computing time and number of nodes of the search tree (RRT) constructed at each iteration.

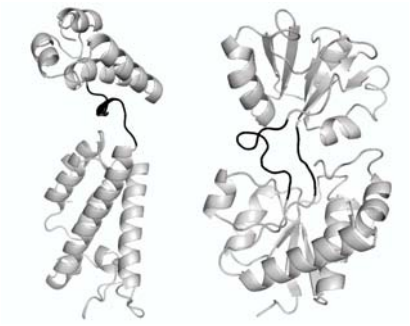

Figure 10: Structures of EIA (left) and LAO (right). The linkers between the different domains are represented in black. The two domains of EIA are connected by a single linker, while two linkers connect the two domains of LAO.

Figure 11: Superposition of the open ADK conformer (black) and the final conformation of the computed transition pathway from the closed conformer (gray).

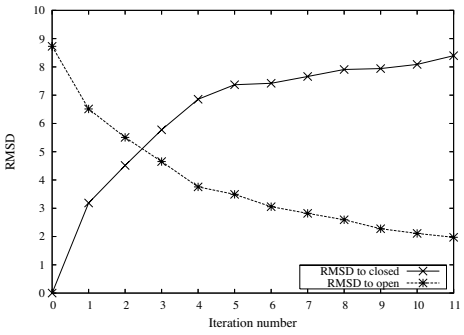

Figure 12: Variation of the RMSD along the computed pathway from the closed conformer toward the open conformer of ADK.

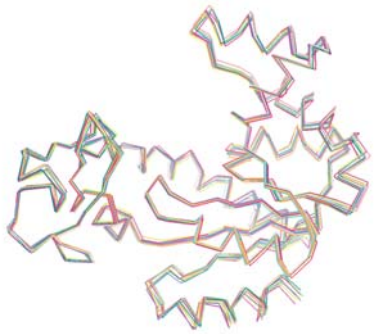

Figure 13: Superposition of the obtained structures for 10 different runs of the first iteration of the NMA-guided RRT search algorithm (Algorithm 2) from the open ADK conformer toward the closed conformer.

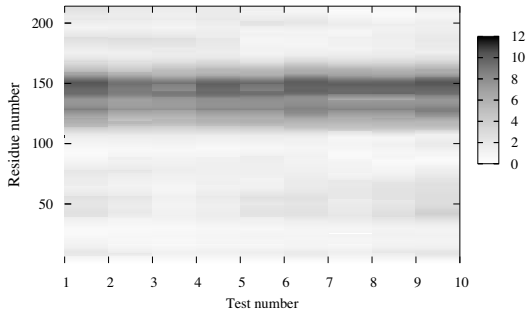

Figure 14: Displacement of the residues for 10 different runs of the first iteration of the NMA-guided RRT search algorithm (Algorithm 2) from the open ADK conformer toward the closed conformer.

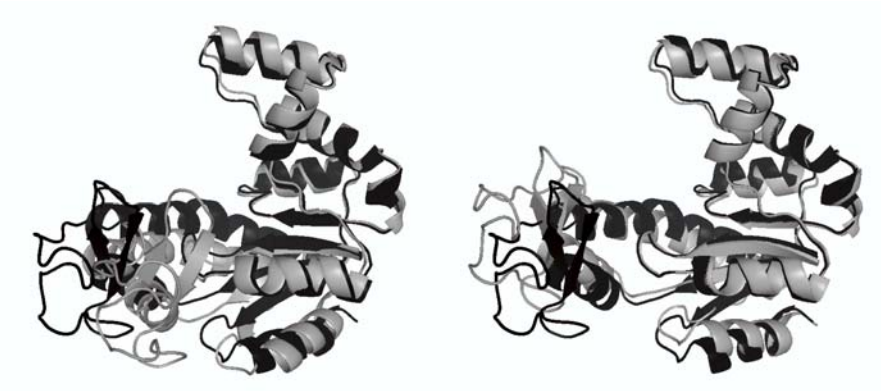

Figure 15: Superposition of the open ADK conformer (black) and the two farthest conformations (gray) obtained after a single iteration of the NMA-guided RRT search algorithm (Algorithm 2) with no bias toward a goal conformation. The region with a more significant motion corresponds to the LID domain. One of the conformations (left) shows a clear tendency toward the closed conformer. The other (right) shows a more open conformation of the LID domain.