Holographic multiscale method used with non-biased atomistic forcefields for simulation of large transformations in protein

2012 J. Phys.: Conf. Ser. 341 012015

(http://iopscience.iop.org/1742-6596/341/1/012015)

Holographic multiscale method used with non-biased atomistic forcefields for simulation of large transformations in protein

L Dupuis[1] and N Mousseau[2]

[1] Département de Biochimie et Centre Robert-Cedergren, Université de Montréal, C.P. 6128, Succursale Centre-Ville,
Montréal, Québec, Canada H3C 3J7
[2] Département de Physique, Regroupement Québécois sur les Matériaux de Pointe (RQMP) et Centre Robert-Cedergren, Université de Montréal, C.P. 6128, Succursale Centre-Ville,
Montréal, Québec, Canada H3C 3J7

E-mail: lilianne.dupuis@UMontreal.CA

E-mail: normand.mousseau@UMontreal.CA

Abstract.

We present a multiscale approach for simulating protein flexibility. The originality of our method is its ability to perform dynamic multiscaling based on continuous revaluation of overlapping areas. The holographic multiscale method overcomes the limitations of motions determined by predefined and fixed high-level descriptions and allows the reproduction of residue-specific impact on large scale motion. The method is tested with two different non-biased all atom implicit solvent forcefields. These show stretched proteins A and G, with maintained secondary structure, folding back near native states in a small number of transition events, demonstrating the advantages of this multiscale approach.

1. Introduction

The success of macromolecular computer simulations rests on describing appropriately two crucial aspects of these complex systems: the interactions between particles, provided by forcefields of various descriptive levels, and the physical geometrical realization of the structure, which can be extracted from experimental data, such as X-ray diffraction or NMR, or constructed by methods for sampling the system's configurational space. In spite of ever growing computational resources, it is still necessary to find the simplest level of description and sampling that will provide the best balance for answering the problem we are interested in.

This right balance is particularly difficult to find when looking at large-scale protein motions, that are often controlled by atomic-scale details but occur on time scale beyond current's general sampling methods. To overcome these opposite requirements, it is necessary to consider multiscaling approaches that can focus on crucial aspects of both of interactions and motions.

Traditional molecular dynamics, for example, describes exactly, within the forcefield limits, motion at all scales, from thermal vibrations to docking and folding. However, because no bias is provided to favor large motion, the large majority of the computing efforts go to describing thermal oscillations, providing little new understanding. Monte Carlo approaches benefit from

considerably more freedom in selecting the appropriate level of motion [1, 2, 3]. However, since these do not generally take forces into account, the size of moves is typically limited by collisions and steric constraints. These are avoided by methods such as Normal Mode Analysis (NMA) [4, 5, 6, 7], a method that identifies directly the softest directions of change by first mapping the protein as an elastic network to compute the lowest-frequency modes of the dynamical matrix, a remarkably good approximation. As low frequencies only depends on the general shape of the molecule [8], NMA can be used with simple coarse-grained level representations [9, 10, 11, 12].

There exists many problems and phenomena, however, where large scale motion of a molecule is affected by mutation of single residues [13, 14, 15, 16] and that cannot be described by coarse-grained harmonic state analysis [17]. For such movements, it is essential to adopt a multilevel representation that adapts to the role of the various molecular structures and incorporates non-harmonic motions. The latter can be achieved with activated methods that propose a non-biased approach for the study of complex transformations affecting the general shape of proteins. These approaches perform continuous evaluations of force curvatures in order to discover energetically feasible transition passages, providing a natural way to leave local minima and identify physically-possible pathways [18, 19, 20, 21]. Focusing on the atomic scale, however, these methods rapidly become dominated by small, local rearrangements that contribute very little to overall configurational changes.

Here, we propose a multiscale algorithm that allows the activated method to focus on large scale motion while retaining the capacity to adapt this representation to atomistic constraints and motion. The holographic multiscale method (HMM) introduces a dynamical definition of the various representation scale that increases considerably the versatility of this approach compared with standard multiscale methods [3, 9, 10, 22, 23]. More precisely, the HMM combines the activation-relaxation technique (ART nouveau), an activated method that has been used extensively for protein folding and aggregation, both in cartesian [20, 24, 25, 26] and in internal coordinates [27], with an adaptive structure-dependent multiscale representation that focuses on large cooperative ensemble motions while allowing, at every step, atomistic relaxation that fully incorporates the role of specific residues and interactions on large scale dynamics.

The structure of this paper is as follows. We begin with a brief overview of multilevel characteristic of proteins, on which the holographic view is based. We then present the details of the holographic multiscale method, including the multiscale projection into reduced internal coordinates. We briefly discuss our in-house implicit-solvent reduced-representation extended OPEP forcefield and then present a few test applications that demonstrate the strength of the algorithm.

2. Methodology

2.1. Basic protein motion

A protein is a chain of a few tens to many thousands of amino acids characterized by a common part composed of a short sequence of a nitrogen and a carbon atom positioned on each side of a central carbon (Cα) and by a lateral chain, or residue, of varying length and composition, rooted at the Cα. Amino acids are linked by their common part through a series of covalent bonds called peptidic links, which form the protein's main-chain or backbone. The presence of covalent bonds restricts effectively the possible motion of each amino acid with respect to its neighbors to the torsion angles situated on both sides of the Cα. With these degrees of freedom, the main-chain can adopt a number of conformations that determine the protein secondary structures — α-helices, β-sheets, loops, random coils, etc. — which are further packed into a three-dimensional organization called tertiary structure. This well-defined organization is guided by delicate balance of backbone–backbone, side-chain–side-chain, backbone–side-chain and protein-solvent interactions, controlled by hydrogen bonds, hydrophobic, polar and charge

groups.

Large structural changes in the protein necessarily involve large modifications of the main-chain torsion angles, ϕ and ψ, located, respectively, around the N-Cα bond, and the Cα-C bond. For example, significant changes in these angles are required to go from $\phi \approx -55$ and $\psi \approx -45$ degrees for α-helices to $\phi \approx -130$ to -160 and $\psi \approx 130$ to 160 degrees for β-sheets. Even though these angles largely characterize the protein spatial organization, other angles and bond-lengths will adapt slightly to every configuration in order to minimize steric effects and optimize stability. The holographic multiscale method combines these two levels of relaxation. Following other large scale motion techniques [28], the HMM focuses one these angles to generate the activation. Its originality comes from the fact that allows all degrees of freedom to adapt during the relaxation phase.

2.2. Overview of the holographic multiscale method

To focus on large scale cooperative motion, the holographic multiscale method proposes a multilevel approach that starts with the protein in a local energy minimum and goes as follows:

(i) The protein conformation is described on the basis of a list of flexible Cα pivots. It is at this moment that a multiscale description is introduced, by the identification of flexible (or *active*) regions, typically associated with loops and random-coil conformations, and rigid regions, those with stable secondary structures. At each active Cα pivot the protein is split into three blocks: the chain fragment preceding the pivot, that following it, and the lateral chain. Each pivot possesses six degrees of freedom associated with the rotation of the last two sections with respect to the first one and to each other. After each displacement step, the all-atom Cartesian space representation of the molecule is regenerated in order to compute the atomistic forces which are projected back onto each overlapping block definition delimited by the various freed pivots (Fig. 3). This generates a smoothened configurational subspace, eliminating small, but non-diffusive, barriers.

(ii) A first activated event is generated, using ART nouveau, on this reduced configurational subspace, leading to a first-order saddle point associated with large-scale cooperative motion.

(iii) From this saddle-point associated with a transition state, the protein is brought into a local-energy minimum using damped molecular dynamics applied to the all-atom Cartesian representation with all degrees of freedom allowed to move.

These three steps represent an ART event, going from one local energy minimum to another, passing through an activated state and following a physically-possible pathway, that allows an efficient sampling of the configurational energy landscape. The next sections describe in more details each of these steps.

2.3. Construction of elastic blocks and projected flexible regions

As mentioned in the previous section, the crucial step in HMM is the construction of the multiscale representation. For this, we first construct a spherical representation of positions and forces projected onto the peptidic plane angular degrees of freedom that determine the rotation of the ϕ and ψ angles. For each of these planes, we compute an orthogonal basis, using the Cα to Cα axis direction as first axis,

$$\overrightarrow{a_1} = \frac{\overrightarrow{r}_{C\alpha 2} - \overrightarrow{r}_{C\alpha 1}}{\| \overrightarrow{r}_{C\alpha 2} - \overrightarrow{r}_{C\alpha 1} \|}, \tag{1}$$

and the perpendicular direction toward the main-chain C-bound oxygen as second axis,

$$\overrightarrow{a_2} = \frac{\overrightarrow{r}_{oxy-C\alpha_1} - (\overrightarrow{r}_{oxy-C\alpha_1} \times \overrightarrow{a_1}) \times \overrightarrow{a_1}}{\| \overrightarrow{r}_{oxy-C\alpha_1} - (\overrightarrow{r}_{oxy-C\alpha_1} \times \overrightarrow{a_1}) \times \overrightarrow{a_1} \|}. \tag{2}$$

The third axis is deduced from a right hand rule

$$\overrightarrow{a_3} = \frac{\overrightarrow{a_1} \times \overrightarrow{a_2}}{\|\overrightarrow{a_1} \times \overrightarrow{a_2}\|}. \tag{3}$$

This forms the absolute base A for each peptidic plane δ. We also need to create its relative base B_δ, which is orthogonal:

$$B_\delta = (A_{\delta-1} \centerdot {A_\delta}^T)^T \quad \text{giving} \quad \overrightarrow{b_1}, \overrightarrow{b_2}, \overrightarrow{b_3} \tag{4}$$

This relative representation allows use to evaluate the angular position (swiveling) of the following area relatively to the preceding, using two spherical angles: the longitude γ and the latitude λ of the Cα-Cα axis (b_1), relatively to the preceding plane, defined as:

$$\lambda = -\arccos \frac{b_{1,2}}{\|\overrightarrow{b_1}\|} \tag{5}$$

and

$$\gamma = \frac{b_{1,3}}{|b_{1,3}|} \arccos \frac{b_{1,1}}{\|\overrightarrow{b_1}\| \times |\sin \lambda|}. \tag{6}$$

Because the blocks are tree-dimensional, a third angle represents the rotation θ of the plane around its own Cα-Cα axis (b_1). As this axis also swivels in space, the evaluation of its rotation requires a comparison with defined referential which we take to be the x axis of the preceding base. We first determine the axis of rotation c between axis b_1 and the x axis:

$$\overrightarrow{c} = \overrightarrow{x} \otimes \overrightarrow{b_1} = (0, b_{1,3}, -b_{1,2}). \tag{7}$$

We then determine the rotation matrix C around this c axis for the angular course needed between axis b_1 and x axis. We obtain this angle from the arc-cosinus of the scalar product between axis b_1 and x axis. This rotation matrix C is then used to apply a theoretical swivelling of the current matrix B toward the x axis of the referential previous block, giving the matrix S:

$$S = (C \centerdot B^T)^T. \tag{8}$$

The first axis of matrix S coincide with the x axis. The second axis of matrix S is used to determine the current rotation angle of the block, using its angular position relative to the y axis:

$$\theta = \frac{s_{2,3} - s_{2,1}}{|s_{2,3} - s_{2,1}|} \arccos \frac{s_{2,2}}{\|\overrightarrow{s_2}\|} \tag{9}$$

The modification of these tree angles corresponds to a holographic swiveling/rotation around the peptidic junction δ-1,δ. With this representation, the entire protein following a Cα pivot is allowed to swivel and rotate relatively to the preceding part. This representation uses 3 angles instead of the 2 torsion ϕ and ψ angles. This means that some variation in the valence angles is allowed here, in addition of the easier dihedral rotations. The valence angles around the Cα may be stressed at the beginning of the activation, but the convergence method will soon relaxed them and conducted the motions to rotate according to phi and psi dihedral angles. The position of the lateral chain is also express in spherical representation relatively to the preceding peptidic plane, including a rotation around its own Cα-Cβ axis.

Atomic positions must be retrieved at each step of the simulation in order to compute the forces. We first get back axis b_1 of the relative base B:

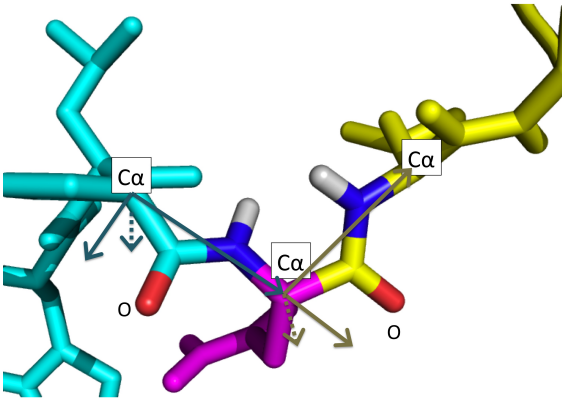

Figure 1. A Cα is present at the root of the residue represented in magenta. An orthogonal basis is defined for the two peptidic planes on each side of this pivot, using the two ending Cα of the plane and the perpendicular direction within the plane, on the oxygen side. We obtain the absolute bases matrices $A_{\delta-1}$ and A_δ. Relative matrix B_δ is then defined as the representation of matrix A_δ in the orthogonal basis of matrix $A_{\delta-1}$. By projecting the cartesian-defined forces on these bases, we obtain spherical rotating/swivelling forces of the whole cyan area relatively to the whole yellow area. The side-chain represented here in magenta is also evaluated relatively to the two large areas, and also simulated as a rotating/swiveling block.

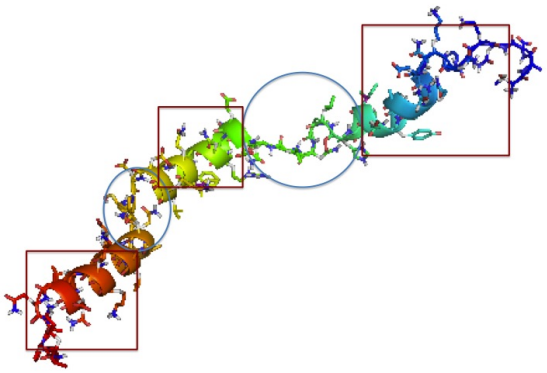

Figure 2. All Cα's of a protein may be defined as pivots. For multiscale simulations, we consider rather some part of the molecule (squares) as structurally stable and define pivots only on Cα's in the flexible zones (circles). Protein A, for example, is defined in the following way: α-helices are considered stable zones and all Cα's in the loops and at the end of α-helices parts bordering them are fully flexible.

$$\overrightarrow{b_1} = (\sin\lambda \times \cos\gamma, \cos\lambda, \sin\lambda \times \sin\gamma). \tag{10}$$

The b_2 axis is obtained using the new θ angle, reversing the steps of equations 7,8,9. From matrix B we get back the absolute base A:

$$A_\delta = ({A^T}_{\delta-1}, {B_\delta}^T)^T. \tag{11}$$

Then the new positions are computed from the initial positions, using the initial absolute base A and the current absolute base A' as transformation matrices.

$$\overrightarrow{r_i}' = ({A'}^T \,.\, A)^T \,.\, \overrightarrow{r_i}, \tag{12}$$

where i is the atomic number and $\vec{r}_i$ is a three-dimensional cartesian position vector.

The rotation forces around the different axes are in correspondence with angles around these axes. The force contribution of all the atoms of the molecule is evaluated for each pair of holographic areas. The lateral chain positioned at the concerned Cα pivot is also designed as a block that ca rotate relatively to them. The contribution of each atom for the block rotation

around a particular axis is inversely proportional to its radius distance to the axis of rotation, and the part of the force that is perpendicular to both the axis and the radius between the axis and the atom.

$$F_{a_l} = \frac{\sum(\overrightarrow{f} \times (\overrightarrow{r} - (\overrightarrow{r} \cdot a_l) \cdot a_l))}{\sum \| \overrightarrow{r} \|} \tag{13}$$

where $l = 1, 2, 3$.

Using the current array of relative angular positions of each overlapping block pair, and the newly computed corresponding arrays of sperical block forces, ART nouveau gives new arrays of relative angular positions. At each step this new array is used to compute back cartesian atomistic positions and then the cartesian atomistic forces from forcefield. Those are used by HMM to update of the spherical holographic block forces used by ART. Fig. 3 summarizes this circular process.

2.4. Sampling method

We use the activation-relaxation technique, ART nouveau [19, 24], to generate energetically favorable transition events from one conformation to the other. Since the method has been reviewed recently [29, 30], we present here only the general scheme as applied into HMM. Starting from a configuration in a local minimum, a random direction of deformation is selected among the available degrees of freedom (pivots). The configuration is slowly deformed. At each step, the lowest eigenvalues of the Hessian matrix are computed using a Lanczos scheme. This random deformation is continued until the lowest eigenvalue becomes negative, indicating the presence of a nearby first-order saddle point. The configuration is then pushed, in the reduced holographic projection, along the eigenvector corresponding to this eigenvalue while the energy is relaxed in the perpendicular hyperplane. This activation phase stops when the total force becomes close to zero and the configuration has converged onto the transition state. During this first phase, motion can take place around all free pivots; the rest of the protein is moved as blocks. For the relaxation phase, the configuration is then nudged over the saddle point and the system is relaxed into a new minimum using, now, the full set of degrees of freedom in cartesian coordinates. The new configuration is then accepted or rejected using a Metropolis criterion.

The advantage of this approach is that, by using the system's forces to direct the activation, collisions are naturally avoided, allowing large collective motions. An example of an ART nouveau step can be seen in Fig. 4 and is discussed in Section 2.6.

2.5. Forcefields

The holographic multiscale method can be used with any non-biased implicit solvation atomistic force field. We present here results with two different force fields: CHARMM19 with the solvation model EEF1 as well as the extended optimized-potentiel for extended proteins (EOPEP) forcefield, an extended version of the OPEP coarse grain forcefield.

For better integration with HMM, we developed an in-house version of CHARMM19, based on Ref. [31]. Following Ponder's approach with Tinker [32], we use our own torsion angle description, based on OPEP's [33], but adapting the prefactors to those of CHARMM19. We combined this potential with the EEF1 solvation model volume exclusion terms [34]. For the latter, we applied ionic neutralization of the side-chains, the N-terminal and the C-terminal, as specified in EEF1.

To distinguish between forcefield and methodological limitations, we also developed our own extended version of the OPEP forcefield. The original OPEP is an implicit solvent reduced-representation forcefield that includes all heavy main-chain atoms as well as a single bead for the lateral chains, with statistically-derived interactions. Among its original features is a four-body interaction term to describe the cooperativity observed in the H-bond formation for stabilizing

secondary structure. The quality of this potential has been well-tested in protein folding and protein aggregation simulations for chains of 70 residues or less [33, 27, 35, 20]. While this potential performs very well for secondary and tertiary structure, it fails to describe precisely more flexible regions where long residues often dominate. EOPEP inherits OPEP's ability to stabilize secondary structure using a cooperativity term for the evaluation of hydrogen bridges as well as provide side-chain hydrogen-bridge. In addition to these terms, EOPEP develops fully the lateral chains to ensure better packing for large residues.

More precisely, EOPEP represents all side-chain's carbon, oxygen, nitrogen, sulfur, and polar hydrogen (linked to O or N) atoms. The various carbon configurations are defined with implicit hydrogens. For these atoms, we adopt the van der Waals atomic dimensions of CHARMM19 [31]. Following OPEP, attractive and repulsive interactions between side-chain atoms are described using a screened Lennard Jones 6-12 term, except for the side-chain hydrogen bridges, for which we use OPEP's Lennard Jones 10-12 term used in the N-H–O interactions. This choice ensures that EOPEP can be used in conjunction with OPEP, providing a multilevel description inside a single protein and decreasing significantly the computational costs while adding the appropriate degree of complexity where needed. Because of their small size, however, proteins studied here are described fully by EOPEP.

Table 1. EOPEP parameters for side-chain atoms. These parameters are used in addition to the covalent bonding and main-chain main-chain interactions already defined in OPEP. EOPEP parameters were adjusted according to the procedure described in the text.

Interaction	Implementation	Partial contribution
Side-chain Hydrogen Bridge		
Carboxyl ogygen	3 terms LJ10-12	As OPEP Main-chain O
Other oxygen receptor	3 terms LJ10-12	0.10 vs Main-chain O
Nitrogen and oxygen donor	3 terms LJ10-12	0.20 vs Main-chain N
Side-chain non polar Carbons	2 terms LJ6-12	$0.05(\mathrm{Kcal/mol})^2$
Sulfur	2 terms LJ6-12	As OPEP MET
Non HB interaction from polar atoms		
Oxygen and Nitrogen	2 terms LJ6-12	0.5 of OPEP ASP
Hygrogen	2 terms LJ6-12	0.05 of OPEP ARG

The value for the extended set of parameters is shown in Table 1. To avoid overstabilizing side-chain–side-chain interactions, the hydrogen bridge strength for lateral chain is considerably reduced compared to the main-chain H-bridge. The hydrogen bridge receptor of carboxyl groups (aspartate, glutamate) conserves main-chain interaction strength, while other oxygen receptor see their contribution decreased by a factor ten, and the contribution of side-chain hydrogen bridge donor (nitrogens or some oxygens) by a factor of 20. These choices reproduce the aqueous immersion effect on them, as such contacts are generally only established consecutively to tertiary structure formation, which are directed by hydrophobicity [36]. Hydrophobic side-chain interactions also follow OPEP and are described by a uniform Lennard Jones 6-12 function, with a 0.05 $(\mathrm{kcal/mol})^2$ well-depth contribution for any type of non-polar side chain carbon atom. Sulfur atoms, for their part, are directly copied from the OPEP methionin behavior.

The EOPEP forcefield was calibrated through extensive stability tests of the native state of protein A as well as calmodulin C and N terminals using long holographic ART simulations.

EOPEP was subsequently used unchanged in folding simulations of these three sequences as well as Troponin C NT and Protein G.

2.6. Description of a typical holographic ART transition event

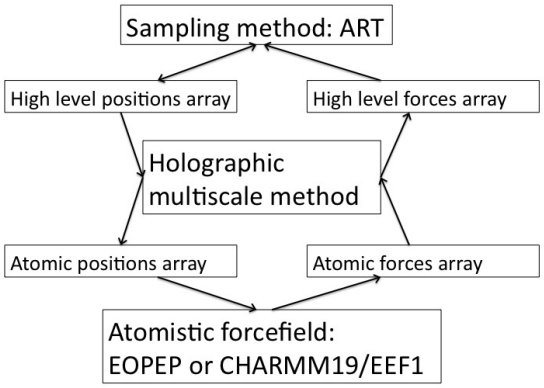

Figure 3. Software organization. The activation and saddle passage steps of ART use the high-level pivot and blocks representation. After each iteration, ART returns new relative angular positions for the various blocks. These are translated back into Cartesian coordinates and the respective force is computed and then projected onto the reduced space corresponding to the high-level description.

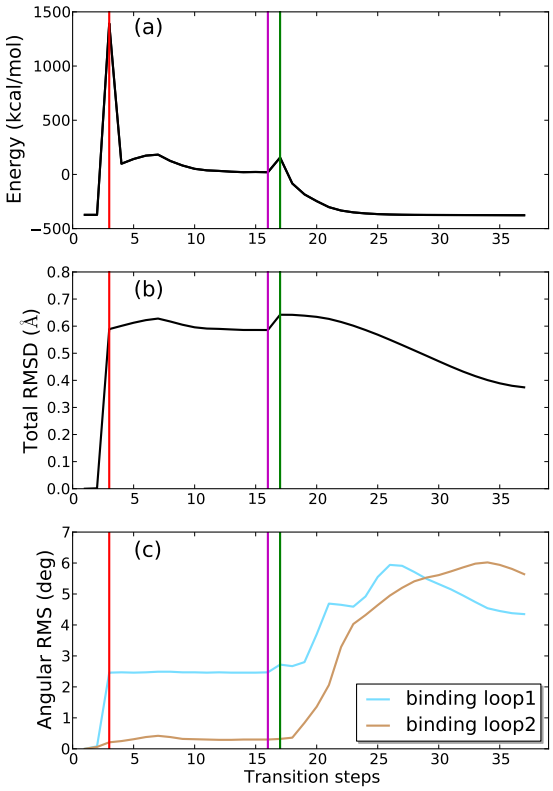

Figure 4. A typical HMM/ART event during an EOPEP folding simulation of Calmodulin NT. (a) Evolution of the total energy as a function of iteration number. (b) Global Cα RMSD measured from iteration zero. (c) Root-mean square angular displacement computed over ϕ and ψ, as measured from iteration zero and averaged over the two binding loops. The open CAM protein model is taken from pdb 1CLL, with residues 1-3 added, totalizing 76 residues. Vertical lines delineates the various phases of an activation/relaxation event. From left to right: push outside of the harmonic basin, convergence to the saddle point, push over the saddle, relaxation into the final minimum.

A typical HMM-ART event takes a few minutes of on a single CPU. Fig. 4 reports details for a typical accepted event during a closing simulation of the open Calmodulin NT model, which undergoes transformation in its binding loops when deprived of its calcium ions [37, 38]. For

this simulation, we defined as pivots residues 18 to 31, which includes the first binding loop, and residues 54 to 67, which includes the second binding loop. The event was started from a energy minimum at -372.9 kcal/mol. The ART nouveau steps are identified on the plot and correspond to the exit from the harmonic basin, the convergence to the saddle point, the push over the saddle and the relaxation to a new minimum. Looking at panels (a) and (b), we note that the energy goes up rapidly as the protein is deformed, reaching 1450 kcal/mol, at the exit of the harmonic basin in a random direction (step 3). Since the convergence to the saddle point is accompanied by a relaxation of the N-1 other excited degrees of freedom, the energy falls rapidly during this next phase but the saddle point is still found at 528.6 kcal/mol above the initial minimum (step 17). This unphysically large energy is due to the fact that only angles around pivots are allowed to move, creating considerable strain in the other degrees of freedom, such as bond lengths and various angles. Indeed, most of this strain is released during the first steps of the relaxation to the final minimum, which takes place in Cartesian space, on all coordinates, and the final structure (step 37) is at 0.4 Å away and 4 kcal/mol (with a total energy of -376.7 kcal/mol) below the corresponding initial state values.

Panel (c) of Fig. 4 reports the root mean square average of ϕ and ψ Ramachandran angles within each binding loop,

$$\theta_{\mathrm{RMS}} = \sum_i \sqrt{(\triangle\phi_i)^2 + (\triangle\psi_i)^2} \tag{14}$$

where i runs on the active residues of loop 1 or 2.

The initial activation takes place mostly in the first binding loop, with relatively little deformation in the second loop. Interestingly, most of the angular displacement in this phase takes place during the exit of the harmonic basin and only small angular adjustments are observed during the convergence to the saddle point. Relaxation allows the full protein to react to this transformation and we see the second loop adjusting significantly during relaxation, creating a large cooperative event that changes considerably the protein conformation.

Table 2 reports the averaged RMSD between initial and final states of events for various simulations. Events can be very large, typically between 2.0 and 4.0 Å. The largest moves are unlikely to lead to low energy structures, however, and at low Metropolis temperature, most moves tend to be rejected. Statistics for accepted events at 300 K show that most lead to displacements are around 1 Å, which is still considerable.

Table 2. Mean RMSD between events: statistics for a few simulation examples

Protein model	Forcefield	Sim	Mean RMSD at saddle all events	Mean RMSD at minima all events	at saddle accepted only	at minima accepted only
Protein A	CHARMM19	5	2.29	2.39	1.27	1.28
Protein A	CHARMM19	16	2.93	2.96	0.96	0.77
Protein A	EOPEP	1	4.13	4.38	1.11	1.03
Protein A	EOPEP	20	3.10	3.25	1.16	1.09
Protein G	CHARMM19	2	2.24	2.03	1.26	0.98
Protein G	CHARMM19	11	2.07	1.94	1.12	0.95
Protein G	EOPEP	2	2.74	2.66	1.25	0.83
Protein G	EOPEP	14	1.93	1.52	1.19	0.69

3. Protein models

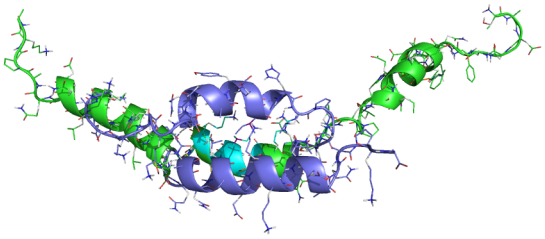

Figure 5. Extended protein A with 3 helices aligned compared to native model. In cyan: the central helix 2. In blue: helices 1 and 3 folded toward helix 2 in the native form. In green: helices 1 and 3 extended away from helix 2, at the start of our simulations.

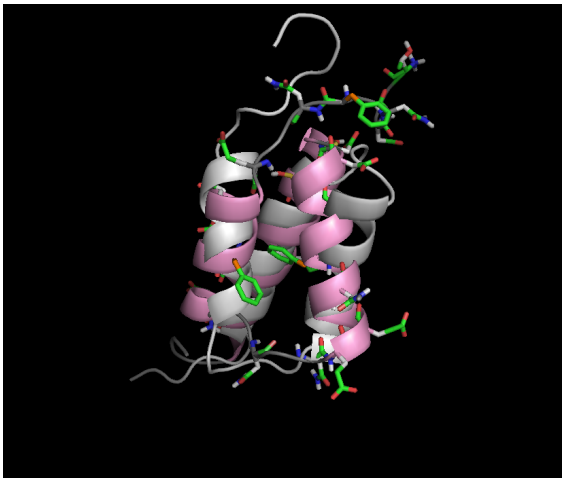

Figure 6. Pink: protein A at end of EOPEP simulation 20. White: the native protein A model.

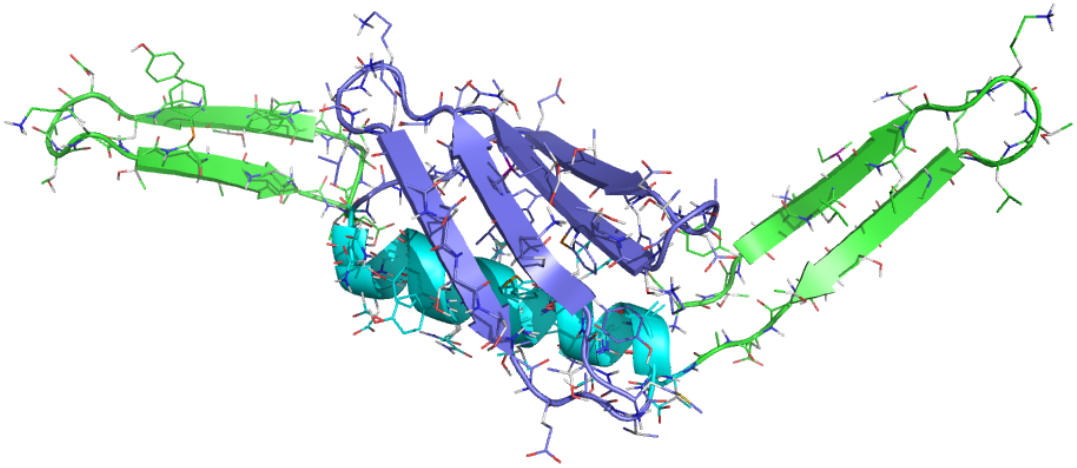

Figure 7. Protein G with beta sheet widely opened, compared to native model. In cyan: the central helix. In blue: the beta sheet formation folded along the central helix in the native form. In green: in our simulations starting model, the beta sheet formation is split away at each end of the central helix

We tested here the ability of the method to perform long distance transformations. To do that, we first stretched two proteins while preserving intact their secondary structure. Folding simulations for protein A domain of protein A (1BDC pdb file), a 60-residue three-α-helix bundle, was started from a configuration with the three helices aligned and a 16.6 Å RMSD with respect to the native configuration (Fig. 5). A 56-residues fragment of protein G model (pdb file 1GB1), composed a α-helix and a four-stranded β-sheet, was also artificially opened to provide a starting conformation at 19.6 Å RMSD away from the native conformation (Fig. 7).

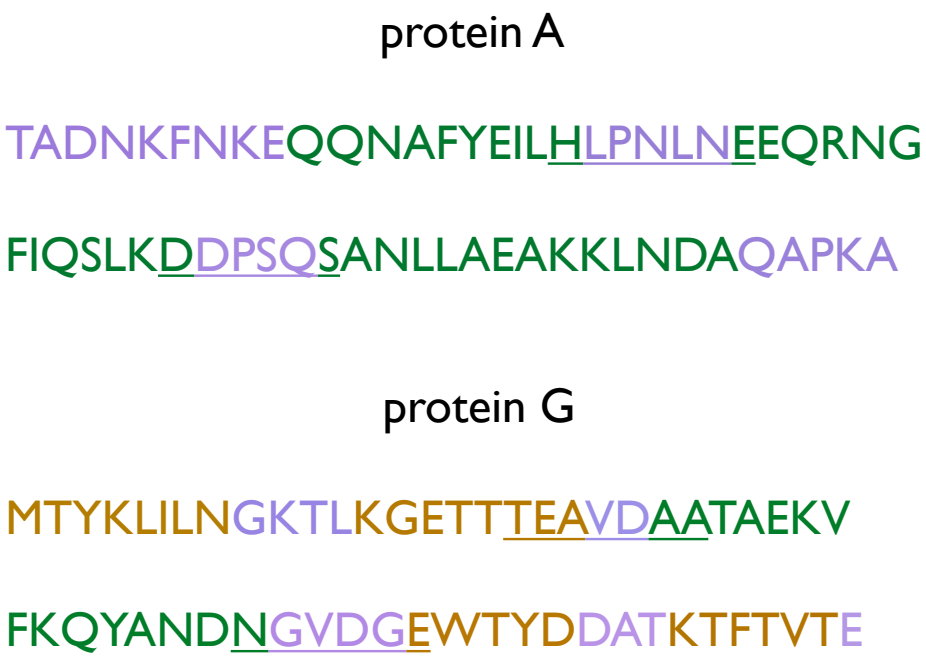

Figure 8. Amino acid sequences of the protein A and protein G domains used in our simulations. α-helices are represented in green, β-strands in orange and the loops in light purple. In the present study, activations and saddle passages were allowed for the Cα pivots underlined in this figure. Relaxations were performed in all-atom mode on the entire sequences.

Figure 2 shows that it is possible to consider some part of the molecule as stable regions and avoid activations in these areas. Figure 8 reports the sequence of the protein A and protein G domains used and the Cα pivots that were allowed for activation. At each activation we performed a random choice of two trios of consecutive C-alpha pivots among the ones underlined in Figure 8. We activate at each of these chosen pivots in random directions for the three peptidic plane angles of the main chain and the for the three rotating block angles of the side chain.

4. Results and Discussion

We performed 24 folding simulations starting from the open version of protein A model, each with 150 to 300 accepted events for CHARMM19 and 300 to 500 for EOPEP at a Metropolis temperature of 300 K (except eop16a, which is discussed below). The evolution of the best runs, both in terms of energy and RMSD are reported in Fig. 9. EOPEP simulations 2 and 20 bring the model under 3.5 Å in respectively 3 hrs and 5.5 hrs CPU time for 50 to 75 events. Simulation 20 reaches 2.8 Å at the lowest-energy state of all simulation after 1.6 day CPU time. Convergence is slightly less for CHARMM19/EEF1, where the best simulation (che21a) comes to 3.8 Å with the native state.

We obtain similar results with protein G. Fig. 10 reports a selection of the 48 folding simulations starting from the stretched protein G model, each running for 200 to 400 accepted events at 300 K. The best RMSD descent is obtained with the CHARMM19/EEF1 simulation number 33, reaching 3.7 Å with the native model. This simulation also displays the lowest energy of all 48 runs using the CHARMM19/EEF1 forcefield. Several EOPEP simulations fall

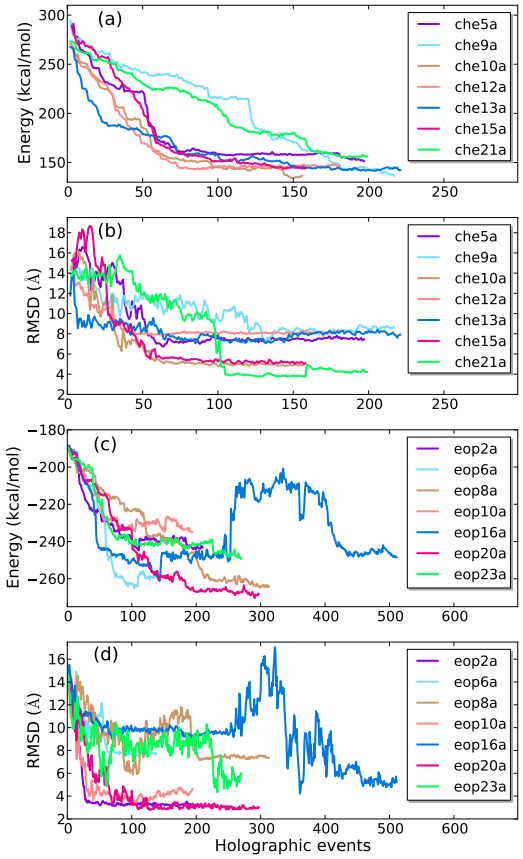

Figure 9. 24 holographic simulations are performed starting from the artificially open model of protein A (three α-helices aligned), using both CHARMM19/EFF1 and EOPEP. From each set we show the four trajectories visiting the lowest-energy structures and the four trajectories visiting structures with the lowest RMSD with respect to the native conformation (only simulation ch5a meets both criteria). (a) Energy and (b) RMSD evolution as a function of the number of accepted ARTn events usign CHARMM1/EEF1. (c) and (d), same but with EOPEP forcefield. The RMSD is measured relative to the Cα of the native form, excluding the unstable ending loops (residues 1-11 and 55-60). All simulations are run at 300 K, except eop16a, which is run at a Metropolis temperature of 4000 K from event 250 to a few events past 400 and then returned at 300 K.

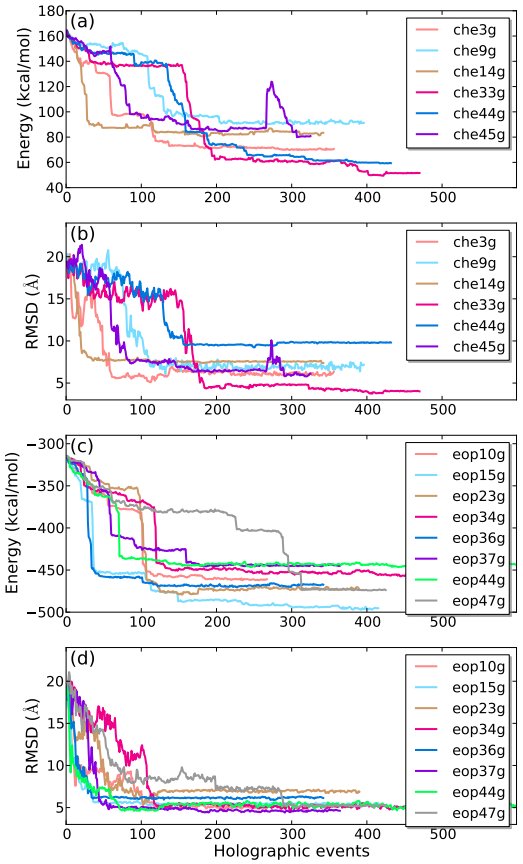

Figure 10. 48 holographic simulations are simulated starting from the artificially open model of protein G. From each set we show the four trajectories visiting the lowest-energy structures and the four trajectories visiting structures with the lowest RMSD with respect to the native conformation (simulations che3g and che33g meet both criteria). (a) Energy and (b) RMSD evolution as a function of the number of accepted ARTn events using CHARMM1/EEF1. (c) and (d), same but with EOPEP forcefield RMSD is relative to all the Cα of the native form. All simulations are run at 300 K, except che45g, which is run at a Metropolis temperature of 4000 K for a few events around event 270 and then returned at 300 K.

below 5 Å RMSD within 7 to 14 hours CPU time.

From HMM point of view, these bests results are excellent as HMM simulations tests of the stability for these two proteins show that the folding runs end up near the effective native state of each potential. Extensive HMM sampling simulations at 300 K starting from the native model for both protein A and protein show a stabilization near 2.5 and 1.7 Å-RMSD for protein A under EOPEP and CHARMM/EEF1 and 2.5 and 3.0 Å-RMSD, respectively, for protein G (data not shown), comparable to the best results of HMM folding.

The fact that only a fraction of all simulations reach the forcefield minima is due to the constant temperature scheme used here. If we simply heat up a structure trapped in a relatively high energy or RMSD basin and allow it to leave this metastable state, it is easy to bring the sequence in more stable state. This is the case for protein A EOPEP simulation number 16 which was trapped in a basin at around 10 Å RMSD after 250 steps. Raising the Metropolis temperature to 4000 K for about 150 events, the run explores configurations up to 17 Å RMSD away from the crystalline state, but rapidly folds to 4.6 Å RMSD after the temperature is brought back to 300 K. A similar but shorter test performed at the end of simulation che45g (protein G with CHARMM19/EEF1) enable the protein to relax from 7 to 6 Å RMSD from the native state.

Comparing the RMSD and energy evolutions in Fig. 9 and 10, we note that the energy drop typically follows the RMSD change, reflecting the existence of periods of local adjustment for minimizing the energy after a large configurational change. While this occurs for the four simulation sets presented in Fig. 9 and 10, the phenomenon is particularly obvious for protein G CHARMM19/EEF1 simulations presented in Fig. 10 (a) and (b): while both energy and RMSD follow a similar drop in value, for each of the 6 presented simulations, the energy relaxation occur systematically tens or more events events after a RMSD abrupt slope.

5. Conclusions and perspectives

This work presents the holographic multiscale method for generating efficiently large range protein motions. Large RMSD and energy steps are generated using the ART nouveau sampling method combined with a multiscale angular representation. In particular, by coupling block motion for the activation to full-scale atomic relaxation, HMM takes full advantages of the secondary structures while allowing it to change in response to large motion. This step is crucial to couple the role of side-chains, for example, with the stabilization of tertiary structure.

Results presented here on two model systems demonstrate the method's efficiency for generating large reorganization along well-controlled pathways. Clearly, however, a description of the energy landscape at the atomistic level increases the competition between local packing and block motion. This makes itdifficult to accept events that get away from local energy basins without destabilizing the whole structure. The use of simulated annealing or other relaxation techniques would greatly help here, and further study will focus on identifying the best algorithm for constructing the proper Metropolis temperature sequence, which could involve an additional set of activated moves on side chains, after each global move, to optimize the local structure before accepting or rejecting the event. The crucial result presented here, however, is that HMM can generate large motions that are not necessarily observed by normal mode analysis but that can nevertheless take place without uncontrolled steric clashes.

In addition to the model systems presented here, HMM has also been applied with success to characterize the folding of two EF-hands proteins, Calmodulin and Troponin C, upon removal of Ca ions. This work, to be published somewhere else, shows that HMM can be used to identify the specific events, atomic interactions and cooperative moves responsible for large scale changes [39].

HMM suffers nevertheless from two limitations. First, the method is not very effective for constrained loops as it requires being able to move a part of a protein with respect to the other,

something that cannot easily be done while ensure a closing constraint. An appropriate selection of the initial deformation in ART nouveau step should be able to lift this inefficiency but more work remain to be done to establish a universal approach.

More complex, however, is the limitation on the forcefield. Because HMM works effectively at zero K, i.e., without thermal vibrations, it must be used with implicit solvent schemes, which are less precise than fully explicit models. The inherent flexibility of implicit-solvent model are obvious, especially with the coupling of the coarse-grained OPEP forcefield with the all-atom EOPEP. More work, however, must be done to ensure that these models work well away from the native state but this is beyond the development of a method such as HMM.

Overall, HMM provides an efficient and effective tool for characterizing the multiscale cooperativity associated with folding and large-scale motion. By providing physically possible pathways, HMM generates ensemble motion that are likely to be close to how real proteins work as long as forcefield are sufficiently reliable.

Acknowledgements

This work was funded in part by the BiT bioinformatic scholarship program, the NSERC, FQRNT and the Canada Research Chair Foundation. Calculations were done using resources from the Calcul Québec.

References

[1] Noguti T Go N 1985 *Biopolymers* **3** 527–546
[2] Fichthorn K A and Weinberg W H 1991 *The Journal of Chemical Physics* **95** 1090–1096
[3] Wells S, Menor S, Hespenheide B and Thorpe M F 1983 *Physical Biology* **2** 127–136
[4] Brooks B and Karplus M 1983 *Proceedings of the National Academy of Sciences* **80** 6571–6575
[5] Levitt M, Sander C and Stern P S 1985 *Journal of Molecular Biology* **181** 423 – 447 ISSN 0022-2836
[6] Janežič D and Brooks B R 1995 *Journal of Computational Chemistry* **16** 1543–1553 ISSN 1096-987X
[7] Tirion M M 1996 *Phys. Rev. Lett.* **77** 1905–1908
[8] Lu M and Ma J 2005 *Biophysical Journal* **89** 2395 – 2401 ISSN 0006-3495
[9] Tama F, Gadea F X, Marques O and Sanejouand Y H 2000 *Proteins: Structure, Function, and Bioinformatics* **41** 1–7 ISSN 1097-0134
[10] Ahmed A and Gohlke H 2006 *Proteins: Structure, Function, and Bioinformatics* **63** 1038–1051 ISSN 1097-0134
[11] Tripathi S and Portman J J 2009 *PNAS* **106** 2104–2109
[12] Zhang Y, Jasnow D and M Z 2008 *PNAS* **104** 18043–18048
[13] Meyer D F, Mabuchi Y and Z Grabarek Z 1996 *The Journal of Biological Chemistry* **271** 11284–11290
[14] Cohen F E and Prusiner S B 1998 *Annual Review of Biochemistry* **67** 793–819
[15] Clarkson M W and Lee A L 2004 *Biochemistry* **43** 12448–12458 pMID: 15449934
[16] Tousignant A and Pelletier J N 2004 *Chemistry and Biology* **11** 1037–1042
[17] Ma J 2005 *Structure* **13** 373 – 380 ISSN 0969-2126
[18] Barkema G T and Mousseau N 1996 *Phys. Rev. Lett.* **77** 4358–4361
[19] Malek R and Mousseau N 2000 *Phys. Rev.* **62** 7723–7728
[20] St-Pierre J F, Mousseau N and Derreumaux P 2008 *J. Chem. Phys.* **128** 0145101–1
[21] Carr J M, Trygubenko S A and Wales D J 2005 *The Journal of Chemical Physics* **122** 234903
[22] Jacobs D J, Rader A J, Kuhn L A and Thorpe M F 2001 *Proteins: Structure, Function, and Bioinformatics* **44** 150–165 ISSN 1097-0134
[23] Mamonova T, Hespenheide B, Straub R, Thorpe M F and Kurnikova M 2005 *Physical Biology* **2** S137
[24] Mousseau N, Derreumaux P, Barkema G T and Malek R 2001 *Mol. Graph. and Modeling* **19** 78–86
[25] Wei G, Derreumaux P and Mousseau N 2004 *The Journal of Chemical Physics* **119** 6403–6406
[26] Santini S, Wei G, Mousseau N and Derreumaux P 2004 *Structure* **12** 1245–1255
[27] Yun M R, Mousseau N and Derreumaux P 2007 *The Journal of Chemical Physics* **126** 105101–105110
[28] Chu J W and Voth G A 2006 *Biophysical Journal* **90** 1572–1582
[29] Marinica M C, Willaime F and Mousseau N 2011 *Phys. Rev. B* **83** 094119
[30] Machado-Charry E, Caliste D, Genovese L, Mousseau N and Pochet P 2011 *J. Chem Phys.*
[31] Neria E, Fischer S and Karplus M 1996 *Journal of Chemical Physics* **105** 1902–1921
[32] Ponder 2010 *http://dasher.wustl.edu/tinker*

[33] Maupetit J, Tuffery P and Derreumaux P 2007 *Proteins* **69** 394–408
[34] Lazaridis T and Karplus M 1999 *PROTEINS: Structure, Function and Genetics* **35** 133–152
[35] Laghaei R, Mousseau N and Wei G 2010 *J. Phys. Chem.* **114** 7071–7077
[36] Baker D 2000 *Nature* **405** 39–43
[37] M Zhang T T and Ikura M 1995 *Nature structural biology* **2** 758–767
[38] Tripathi S and Portman J J 2008 *The Journal of Chemical Physics* **128** 205104
[39] Dupuis L and Mousseau N *in press, J. Chem. Phys.*