BRIEF RESEARCH REPORT
published: 15 October 2019
doi: 10.3389/fmolb.2019.00104

Structural Transition States Explored With Minimalist Coarse Grained Models: Applications to Calmodulin

Francesco Delfino[1,2*†], *Yuri Porozov*[1,3†], *Eugene Stepanov*[4,5], *Gaik Tamazian*[6] and *Valentina Tozzini*[2]

[1] I.M. Sechenov First Moscow State Medical University, Moscow, Russia, [2] Istituto Nanoscienze – CNR and NEST-Scuola Normale Superiore, Pisa, Italy, [3] ITMO University, St. Petersburg, Russia, [4] St. Petersburg Branch of the Steklov Mathematical Institute of the Russian Academy of Sciences, St. Petersburg, Russia, [5] Department of Mathematical Physics, Faculty of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia, [6] Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia

OPEN ACCESS

Edited by:
Gennady Verkhivker,
Chapman University, United States

Reviewed by:
Marc Delarue,
Institut Pasteur, France
Peng Tao,
Southern Methodist University,
United States

***Correspondence:**
Francesco Delfino
delfinofrancesco90@gmail.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 10 May 2019
Accepted: 24 September 2019
Published: 15 October 2019

Citation:
Delfino F, Porozov Y, Stepanov E, Tamazian G and Tozzini V (2019) Structural Transition States Explored With Minimalist Coarse Grained Models: Applications to Calmodulin. Front. Mol. Biosci. 6:104. doi: 10.3389/fmolb.2019.00104

Transitions between different conformational states are ubiquitous in proteins, being involved in signaling, catalysis, and other fundamental activities in cells. However, modeling those processes is extremely difficult, due to the need of efficiently exploring a vast conformational space in order to seek for the actual transition path for systems whose complexity is already high in the stable states. Here we report a strategy that simplifies this task attacking the complexity on several sides. We first apply a minimalist coarse-grained model to Calmodulin, based on an empirical force field with a partial structural bias, to explore the transition paths between the apo-closed state and the Ca-bound open state of the protein. We then select representative structures along the trajectory based on a structural clustering algorithm and build a cleaned-up trajectory with them. We finally compare this trajectory with that produced by the online tool MinActionPath, by minimizing the action integral using a harmonic network model, and with that obtained by the PROMPT morphing method, based on an optimal mass transportation-type approach including physical constraints. The comparison is performed both on the structural and energetic level, using the coarse-grained and the atomistic force fields upon reconstruction. Our analysis indicates that this method returns trajectories capable of exploring intermediate states with physical meaning, retaining a very low computational cost, which can allow systematic and extensive exploration of the multi-stable proteins transition pathways.

Keywords: proteins conformational transitions, classical molecular dynamics, coarse grained models, transition path sampling, minimal action path, PROMPT

INTRODUCTION

Signaling is a core activity in cells. Most of the signaling processes are regulated by bi- (or multi-) stable proteins, which can undergo conformational transitions in response to changes in environmental conditions or stimuli of different origin (Grant et al., 2010). This class includes among others, G-proteins coupled receptors (Weis and Kobilka, 2008) such as Rhodopsins (Tavanti and Tozzini, 2014) and other transducers, e.g., Calmodulin (Wenfei et al., 2014), and a vast number of enzymes undergoing conformational changes during their activity, such as the HIV-1 protease

(Tozzini et al., 2007). The structural variations are usually quite large, therefore atomistic molecular dynamics (MD) simulations might not be the most proper method to address them, because the slow transition kinetics requires simulations exceeding the currently reachable time and space scales. In addition, the atomistic representation with standard force fields (FF) is not warranty of accuracy for the strongly distorted and out of equilibrium transition states (Best and Hummer, 2009).

Strategies to overcome these difficulties involve different actions. On one side, adopting simplified low-resolution descriptions of the system such as coarse-grained (CG) models (Tozzini, 2005) reduces the computational cost and allows performing more efficient sampling of the conformational space. This advantage comes at the cost of increasing the empirical content of the FF, and consequently reducing predictive power and transferability. A compromise between accuracy and predictive power (Tozzini, 2010) is reached by including some *a priori* knowledge of the system, in different forms, such as, e.g., a (partial) bias (Tozzini and McCammon, 2005; Spampinato et al., 2014) toward reference structures. This appears a reasonable compromise especially in the case of the search of the path between two given structure, when the system must in any case be forced to have them as stable states.

On the other side, one can act by simplifying the sampling algorithm, e.g., using morphing related methods (Weiss and Levitt, 2009; Koshevoy et al., 2014; Tamazian et al., 2015) without relying on any specific FF. In particular PROMPT (Koshevoy et al., 2014; Tamazian et al., 2015) employs an approach based on the optimal mass transportation problem including physical constraints of geometric nature (Evans and Gangbo, 1999). Methods based on the action minimization of simplified FFs, such as MinActionPath (Franklin et al., 2007), can be thought as located between the two approaches. The combination of the different sampling methods with the different representations of the systems and its interaction has given rise in the last decades to a huge number of approaches, which has also posed the problems of their comparison and assessment (Seyler et al., 2015).

In this work, we first apply a minimalist CG model for proteins to the test case of Calmodulin, chosen because of its large conformational transition upon calcium binding. We perform molecular dynamics simulations in different conditions to sample the transition path. We then compare these results with those of the simplified path sampling methods.

SYSTEM AND METHODS

The Coarse Grained Model

The coarse graining procedure we consider in this work is schematized in **Figures 1A,B**, reporting the atomistic representation of a protein chain and the minimalist CG (MCG) representation in which only the Cα atoms are present. The choice of Cα as the representative atom of the amino-acid bead allows uniquely representing the secondary structure by the internal variables α, θ (Tozzini et al., 2006). The interactions are described by an empirical FF, derived from an energy potential U with a form similar to the atomistic ones, separated in bonded and non-bonded interactions

$$U = \sum_{\text{bonds}} u_i^b\left(d_i\right) + \sum_{\text{bond angles}} u_i^\theta\left(\theta_i\right) + \sum_{\text{dihedrals}} u_i^\phi\left(\phi_i\right) + \sum_{i>j} u^{nb}\left(r_{ij}\right) \quad (1)$$

d_i, θ_i, φ_i being the bond distances, angles, and dihedrals describing the local geometry of connected beads and r_{ij} distances between non-bonded ones (see **Figure 1B**). The functional forms (reported in **Table 1**) are somewhat more complex than those used in atomistic FFs: while u_i^b are holonomic restrains, the u_i^θ and u_i^ϕ take forms accounting for the anharmonicity of the CG interactions; in addition, the parameters are chosen to account for the different geometrical stiffness of the secondary structures, assigning different values to helices and sheets (see **Table 1**)[1]. The non-bonded interactions occur between couples not already involved in a bond, bond angle or dihedral interaction and are separated in local and non-local part

$$\sum_{i>j} u^{nb}\left(r_{ij}\right) = \sum_{i,j|r_{ij}<r_{cut}} u_{loc}\left(r_{ij}\right) + \sum_{i,j|r_{ij}>r_{cut}} u_{nl}\left(r_{ij}\right) \quad (2)$$

both represented by a Morse potential, with the local term retaining a bias toward a reference structure (see **Table 1**). In this work the local/non-local separation is based on a geometric criterion: all the non-bonded couples whose distance is less than $r_{cut} = 8.5$ Å in the reference structure are considered local, the others are considered non-local. The cutoff value used here was previously shown to include all the relevant H-bonds and other possible specific interactions such as disulfide or salt bridges (Trovato and Tozzini, 2012). The parameters of the Morse potential, were optimized in our previous works including a dependence on r_0 (distance in the reference structure) in order to reproduce stronger interaction in the H-bonding range and weaker ones in the hydrophobic range (Di Fenza et al., 2009) (see **Table 1**). Since here we are not interested in the accurate simulation of the inter-protein interactions, the non-local part is represented by a generic amino-acid independent potential reproducing an average level of hydrophobicity (**Table 1**), instead than with a complex matrix of amino-acid dependent potentials (Trovato et al., 2013).

Simulation Setup and Transition Path Extraction

MD simulations were performed in canonical ensemble using the Langevin (stochastic) thermostat. The timestep was set at 0.01 ps. Simulations had different length, between 20 and 50 ns. The data dumping frequency was on average 0.1 ps^{-1}. Simulations were performed with the two different CG FFs (hereafter FF$_A$ and FF$_B$) generated with a bias toward closed and open states (A and

[1] The continuous dependence of the k$_\theta$ elastic constant is a variant with respect to previous works using step-wise dependences (e.g., Di Fenza et al., 2009), which improves the numerical stability of the model.

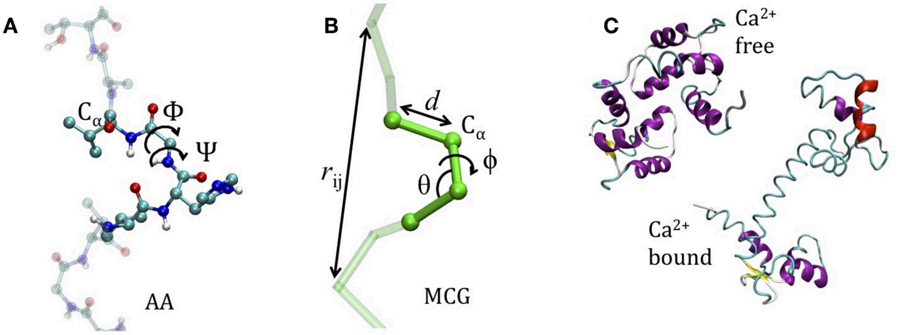


FIGURE 1 | The model system. **(A)** The atomistic representation of the protein chain (side chains are omitted for clarity) **(B)** coarse grained representation. In both cases the internal variables are reported. **(C)** The apo-closed form (named A) and the calcium-bound open form (named B) of Calmodulin (pdb codes: 1WRZ, 1EXR).

B, respectively), and at different temperatures. Simulations were performed with DL_POLY [vs. 4.08 (Bush et al., 2006; Todorov et al., 2006; Boateng and Todorov, 2015)] and the input was generated with proprietary software.

In order to extract a transition path from the trajectory, we first define the parameter σ based on the root mean square deviation ($\mathrm{RMSD}_{\mathrm{A/B}}$) of a configuration $\mathbf{r} = \{x_i, y_i, z_i\}$ from the reference structures $\boldsymbol{r^{A/B}}$[after alignment (Humphrey et al., 1996)[2], to eliminate roto-translations]

$$RMSD_{A/B}(\mathbf{r}) = \sqrt{\frac{1}{N}\sum_i \left(x_i - x_i^{A/B}\right)^2}$$
$$\sigma(\mathbf{r}) = \frac{1}{2}\left(\frac{RMSD_{\ A}(\mathbf{r}) - RMSD_{\ B}(\mathbf{r})}{RMSD_{\ A,B}}\right) + \frac{1}{2} \qquad (3)$$

σ ranges between 0 (in A) and 1 (in B), is a rough measure of the transition advancement. Clearly, structures with the same $\sigma(\boldsymbol{r})$ can have different conformations, with different distances from A and B, accounted for by $\mathrm{RMSD}_A(\mathbf{r})$ and $\mathrm{RMSD}_B(\mathbf{r})$ separately, since the calculation of σ in practice operates a projection of the 2-dimensional path in the $\mathrm{RSMD}_A/\mathrm{RSMD}_B$ plane onto a line connecting A and B. Therefore, the scatter plot $RMSD_B$ vs. $RMSD_A$ will also be considered to have more specific information on the transition path. σ is used to compare the properties of structures with similar transition advancement from the three different methods.

In order to identify a limited number of relevant points along the trajectory, we applied the principal path (PP) clustering algorithm (Ferrarotti et al., 2018) to the MD trajectories and extracted reduced trajectories, which retain the salient properties of the original ones. The PP algorithm is a regularized version of the k-means clustering algorithm (Arthur and Vassilvitskii, 2007), based on the evaluation of a cost functional composed of two parts: the sum of the squared distances of each point from its respective representative structure, and the sum of the squared distances between adjacent representative structures. The relative weight of the two components—the regularization parameter s—is obtained by the Bayesian evidence maximization. The cost functional can be interpreted as an energy, thus the Bayesian posterior probability function is set proportional to the exponential of its negative. The result of the clustering is a "cleaned-up trajectory" of representative structures, used to evaluate σ and energy profiles.

Energies were evaluated both with the CG FFs and at the atomistic level. To this aim, the atomistic structures were rebuilt from the MCG models using Pulchra (Rotkiewicz and Skolnick, 2008) without any local optimization, then explicitly hydrated and locally optimized using the OPLSe (Harder et al., 2016) FF with explicit solvent and the Polak-Ribiere conjugate gradient algorithm (Polak and Ribiere, 1969) keeping the backbone frozen during the minimization. The calculations were performed with Schrodinger 2018-2, MacroModel (2019).

PROMPT and MAP Path Search

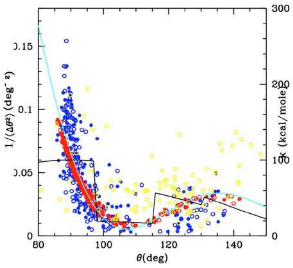
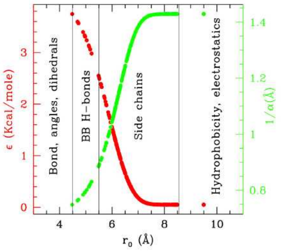
The PP clustering trajectories are compared with the trajectories obtained from other transition analysis methods. The method MinActionPath (Franklin et al., 2007) (MAP) employs differential equations, obtained by minimizing an action functional including a very simplified potential term representing the protein as a network of harmonic interactions (the elastic network model, ENM) (Tirion, 1996). The equilibrium distances are taken from the reference structures, making the ENM the simplest completely biased model. The solutions to the pair of differential equation are merged by requiring continuity between them. The final result is a single trajectory connecting the two states, reproducing the energy profiles of the mono-stable ENMs near A or B, and with a continuous crossover region.

On the other hand, PROMPT (Tamazian et al., 2015) [PRotein cOnformational Motion PredicTion[3]] connects states A and B avoiding relations to any specific FF, by using only structural information. The protein is represented at the CG level and each protein conformation is handled as a set of internal coordinates. The transition path is first guessed e.g., using linear interpolation between extremal configurations $\boldsymbol{r}^{\mathrm{A}}$ and $\boldsymbol{r}^{\mathrm{B}}$. The "admissible motions" are defined, as those preserving all the bond lengths b_i^J and other physical constraints related to

[2] Alignment is performed by means of the built-in extension "RMSD Trajectory Tool" of the graphics software VMD.

[3] Implemented in a publicly available toolbox for MATLAB with its source code on GitHub (http://github.com/gtamazian/PROMPT) and MATLAB File Exchange (http://www.mathworks.com/matlabcentral/fileexchange/49054- prompt).

TABLE 1 | Functional forms (first and second columns) and parameterization (third column) of the MCG FF.

FF term	Functional form	Parameterization
Bond $u_i^b\ (d_i)$	Restrains	d_i from the reference structure (∼3.8 Å)
Bond angle $u_i^\theta\ (\theta_i)$	$\frac{1}{2}k_i^\theta\ \left(\cos\theta - \cos\,\theta_0^{\mathrm{i}}\right)^2$	θ_0^i from the reference structure $k_i^\theta = \frac{k_i}{\sin^2\theta_0^{\mathrm{i}}} \quad k_i = B\left(\frac{\sin\left(\beta\theta_0^{\mathrm{i}}\right)}{\beta\theta_0^i}\right)^2 + k_0$ B = 3,000 Kcal/mole k_0 = 10 Kcal/mole β = 1.667 
Dihedral $u_i^\phi\ (\phi_i)$	$A_i^\phi\left[1 - \cos\left(\phi - \phi_0^{\mathrm{i}}\right)\right]$	ϕ_0^i from the reference structure $A_i^\phi\left[\frac{\mathrm{Kcal}}{\mathrm{mole}}\right] = \begin{cases} 25 \ \ if\ \phi_0 \le 80\mathrm{deg} \ \ \ \mathrm{helices} \\ 5 \ \ \ if\ \phi_0 > 80\mathrm{deg} \ \ \ \mathrm{strands} \end{cases}$
Local $u_{loc}\ (r)$	$\varepsilon^{ij}\left[\left(e^{-\alpha^{ij}\left(r-r_0^{ij}\right)} - 1\right)^2 - 1\right]$	r_{cut} =8.5 Å $\varepsilon^{ij} = 3.8\ e^{-(r_{ij}^0/6.1)^8} + 0.05$ $\alpha^{ij} = 2.2\ e^{-(r_{ij}^0/6.1)^8} +\ 0.70$ 
Non local $u_{nl}\ (r)$	$\varepsilon\left[\left(e^{-\alpha(r-r_0)} - 1\right)^2 - 1\right]$	$r_0 = 9.5$ Å $\varepsilon = 0.05\frac{Kcal}{mole}$ $\alpha = 0.70A^{-\ 1}$

An illustration of the statistics-based parameterization procedure is also reported in the plots. Upper plot: The dots represent the inverse bond angle fluctuations as a function of the bond angle, evaluated using atomistic simulations of different test proteins (yellow a globular protein, blue the calmodulin itself, different symbols for different runs). This curve can be fitted as damped sin (cyan line). Assuming statistical equilibrium one has an angle dependent effective elastic constant from the equation $k' = k_B T / < \theta^2 >$. A further factor $1/sin^2(\theta_0)$ accounts for the non-exactly harmonic functional form used here (i.e., harmonic cosine) leading to the final functional form for k_θ reported in the table, which accounts for the secondary structure dependence of the elastic constant (stronger for helices with $\theta_0 \sim 90°$, softer for strands with $\theta_0 > 110°$). Red dots show the result from a simulation with MGC model with this parameterization. The black line reports the previously used parameter dependence for comparison. For the dihedral term a similar secondary structure dependent parameterization is used, expressed through a simpler step wise dependence on the dihedral value. The non-bonded interactions parameters are reported in the lower plot: dependence of the well depth (ε) and interaction range (1/α) on the equilibrium distance (the shorter the equilibrium distance, the stronger, and shorter ranged the interaction). The plot also reports typical interactions included in the corresponding ranges. In all cases, the 0 subscript indicates the rest value of the corresponding variable. i or i, j apices are the Cα indices (e.g., r_0^{ij} is the rest value of the distance between i and j Cαs).

bond and dihedral angles (i is the index running along the internal coordinate, and J labels the configuration along the path, from A to B). The path connecting A and B is therefore found by minimizing a kinetic only action integral within the space of admissible motions factorized by rigid roto-translations. The infinite-dimensional variational problem is addressed by discretizing the path between A and B and solved by means of the gradient descent method. The admissible motions are searched by changing the internal free variables of the systems, i.e. $\{\theta_i^J, \phi_i^J\}$ in MCG model; θ_i^J is treated by interpolation when possible. The detailed description and formal comparison of the three method is reported elsewhere (Delfino et al., in preparation). Energies along MAP and PROMPT trajectories were compared using both atomistic (upon rebuilding and side chain optimization as already explained) and MCG FFs.

RESULTS

Molecular Dynamics of the Open-Closed Transition of Calmodulin

Calmodulin (Cam) displays two very different conformations (Wenfei et al., 2014), depending on the environmental calcium concentration. The two extremal structures of Cam, i.e., closed (A) and open (B) (see **Figure 1C**), correspond to the apo and Ca^{2+}-bound state, respectively. Because these are, *de facto,* distinct proteins, having different ligands, it is conceptually correct to use two distinct FFs and to perform LD simulation started from A using FF_B to reproduce the A→B transition occurring upon Ca^{2+} binding, and, *vice-versa*, using FF_A for the B→A inverse transition occurring upon Ca^{2+} release. A few data are available for the difference in Gibbs free energy between the folded and denatured proteins ranging between $\Delta G_A \sim 1.5$–3.5 (Masino et al., 2000; Rabl et al., 2002) kcal/mole for the A state and $\Delta G_B \sim 4.5$–6.5 kcal/mole for the B state (Masino et al., 2000). Energy alignment is not straightforward, however, one might assume the denatured state as reference, and infer that B state is more stable than A of about 2–4 kcal/mole.

The A–B transition was simulated with LD, in both senses, at 300 K (RT) and at 130 K (complete simulation data in the **Supplementary Material**). **Figures 2A,B** reports the energies along the LD simulations. In both cases the transitions are clearly visible in the evolution of σ, passing from 0 to 1 (A→B, green) or from 1 to 0 (B→A, red), though they occur at different times, depending on the simulation parameters and on the FF. In particular, the closed to open transition (green) occurs earlier and more directly, while the inverse open to close transition appears to explore an intermediate conformation with σ∼0.4–0.5 for tens of ns before reaching the final state. This is better seen in the RMSD scatter plots reported in panels c and d: the intermediate state, located in the upper right off diagonal part of the plot, persists also after the clustering procedure (joined dots in **Figures 2C,D**) and is present at high and low temperature, although in the low one it is pushed toward the diagonal. It corresponds to a compact globular conformation, favored over the completely open one by hydrophobicity, but in which the specific contacts of the closed conformation are not formed (see the inset in **Figures 2C,D**, red structures). In this work Cam is used only as an example, therefore exploring in detail its transition is out of our scopes. However, we remark that the presence of such mis-folded transition intermediates was previously documented (Wenfei et al., 2014). The intermediate is not visible in the A→B simulations (green), in which the system passes rapidly to B, not even in the PROMPT and MAP trajectories, lying near the diagonal line joining A and B in the RMSD plot. These, additionally, display distorted conformations in the intermediate σ regions. An inspection to the structures with σ∼0.5 (reported in purple and cyan in **Figure 2C**) shows distortions in the central helix and too contracted terminal regions in the PROMPT structure, and broken chain in the MAP structure.

Data Clustering and Comparison With PROMPT and MAP

While MAP and PROMT return transition paths made of a few points, the MD simulations explore a large portion of the conformational space returning thousands of conformations. Therefore, in order to compare the methods, we first performed a post-processing and clean-up of the MD trajectories to select a limited number of representative states along it. This can be done in several ways. **Figure 3A** reports a simple averaging procedure: the structures are first ordered according to their σ value (red and green dotted/dashed lines), so that A→B transition is read from left to right and B→A from right to left. Once again, the formation of an intermediate cluster at $\sigma = 0.4$–0.5 is clearly visible in the B→A simulations, beside the large cluster of A type structures and of B type structures in the A→B simulations, respectively. The structures are then grouped according to their σ value in a given number of regular σ intervals; the average energy evaluated in each interval is reported in the plot, for the A→B (green) and B→A (red) simulations at 300 and 130 K (dots with error bars). Interesting enough, transitions occur in all cases with a gain of ∼20 Kcal/mole (as measured from the starting state, i.e., in each case the opposite of the stable one), irrespective of the temperature and of the FF. As said, comparing the energies resulting from two different FFs is not straightforward. In this case, an inspection of **Figures 2C,D** shows that the simulation trajectories with FF_A and FF_B get particularly near in a region of the $RMSD_A$-$RMSD_B$ plane corresponding to σ∼0.4, indicating that in that area structures belonging to different trajectories are similar. Aligning the energy values for that value of σ in the plot of **Figure 3A** generates a small shift leading to B structure more stable than A one of about 3–4 kcal, roughly corresponding to the experimental evaluation. The resulting "activated state structure" corresponds to the intermediate found in the B→A simulations, which turns out to be located ∼10 Kcal/mole above the A/B states. This "barrier" value seems rather independent on the simulation temperature, whose effect appears to be a rigid shift of the average energies.

While the described procedure gives reasonable values of the energies, representative structures along the trajectories are more properly selected via the PP algorithm. This returns a user-defined (20 in this case) number of elements, which are not elements belonging to the trajectories they represent, but rather elements optimizing the structure variance within the trajectory. As a consequence, the energy profiles obtained evaluating the FF_A and FF_B energies onto them (**Figure 3B**, solid lines and squared symbols) are rather regular and lie lower in energy with respect to parent trajectories, shown by lines connecting circle symbols (obtained selecting the nearest elements to the optimal ones, filled and empty dots connected by dotted and dashed lines). Remarkably, even after post processing, the main features of the simulation remain: the cluster located at σ∼0.4–0.5 is well-represented in FF_A simulations, and is located about 10 Kcal/mole above with respect to A and B states.

The optimal element trajectories extracted from the low temperature runs are also reported in **Figure 3C** to be compared with the energies evaluated from the MAP and PROMPT trajectories using the MCG FFs. Even after a local optimization, the energies from MAP and PROMPT rapidly increase producing a very large energy barrier at intermediate σ values. An inspection of the structures (reported as insets in the plot) reveals that these arise from severe distortion of the backbone (especially for MAP) and/or steric clashes (both). In particular, the high energy of the intermediate from PROMPT seems to be due to steric clashes in one of the two ends of the protein (highlighted with a yellow circle in **Figure 3C**.

Clearly, higher energies on the MAP/PROMPT paths evaluated with MCG FFs are expected, since the low energy path extracted with PP from simulations minimize the MCG Hamiltonian. Therefore, in order to clarify if this energy difference reflects a real larger stability of MCG derived conformations, we rebuilt the atomistic structure of the paths

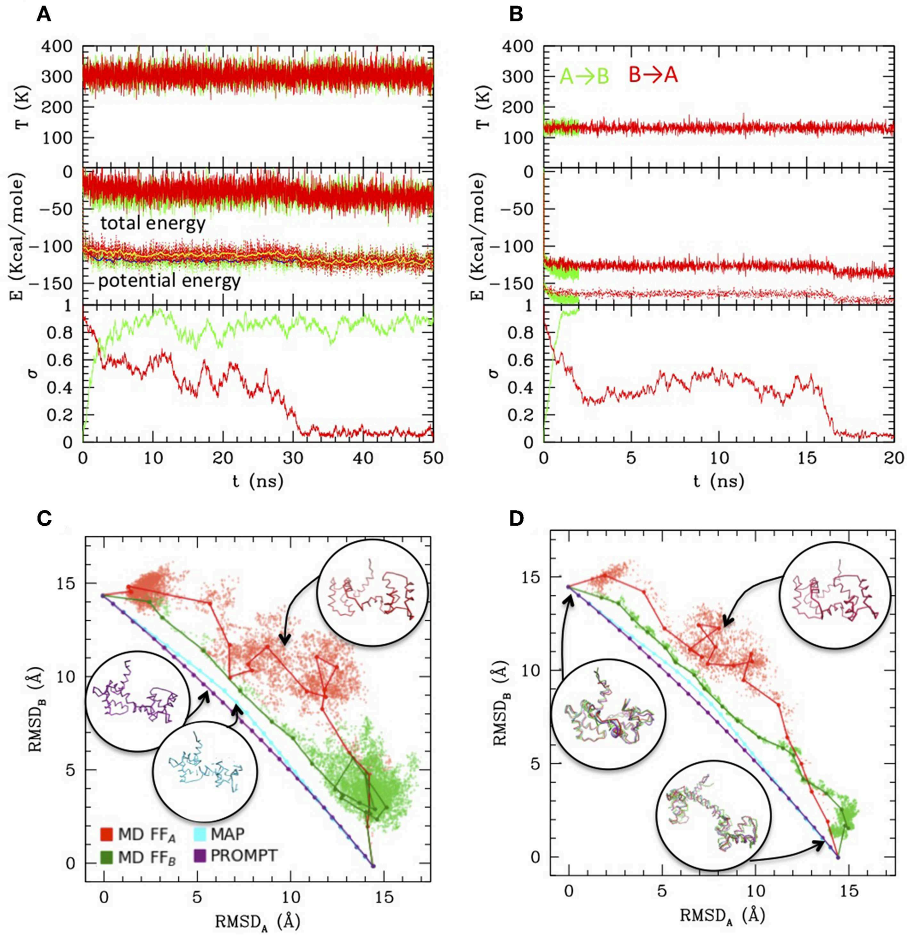


FIGURE 2 | Simulations results from Langevin dynamics at 300 K, $\gamma = 8$ ps^{-1} **(A)** and 130 K, $\gamma = 2$ ps^{-1} **(B)**. Temperature (upper plots), total and potential energies (central plot) and σ are reported along the simulations from A to B (using FF$_B$ and starting from configuration A, green lines), and from B to A (using FF$_A$ and starting from configuration B, red lines). For the 300 K simulation also the running averages are reported for the potential energy as yellow and blue lines, respectively. **(C,D)** Scatter plot of the LD simulations (same color coding as previous) compared with MAP and PROMPT paths evaluation (color coding as in the legend of **C**). The connected dots are the representative elements of the PP clustering procedure. Sample configurations are reported in colors corresponding to the lines and their approximate location in the plots are indicated by arrows.

evaluated with all methods and compared their energies evaluated with the atomistic FFs (**Figure 3D**), after optimization of the side chain conformation keeping fixed the backbone structure. All methods give comparable energies for structures near A and B states, where in some cases PROMPT and MAP seems to work better than MCG models. However, the atomistic analysis confirms the strong instability of MAP derived structures, displaying unphysical backbone conformation, as shown by the reported Ramachandran plot (upper right inset of **Figure 3D**). The instabilities of the PROMPT profile are confirmed in the central σ∼0.2–0.8 region, although the Ramachandran plot (central inset) is regular even in there. In fact, in agreement with what found in the MCG model, the instability is not due to a wrong backbone conformation, but to steric clashes in the highlighted area (yellow circle), displaying two sheets whose relative conformation is too close and not correctly aligned. The complete set of structures and energy data is reported as **Supplementary Material**.

SUMMARY AND CONCLUSIONS

In this work we set up a simulation paradigm for finding the transition path of proteins undergoing large conformational transitions, which is a long-standing problem of biophysics. Proteins are modeled by a Cα based coarse-grained representation, while the transition path is explored via classical molecular dynamics simulations with FFs partially biased toward the reference structures. The selection of a representative trajectory among the huge number of configurations explored during molecular dynamics simulations is accomplished by means of the principal path clustering algorithm, which managed to single out trajectories close to those of minimum free energy, yet capable of exploring intermediate states, with a very low computational cost. The comparison with minimal action path and PROMPT can be summarized as follows: MAP returns structures which are reasonable in the near vicinity of the references states, but is unable

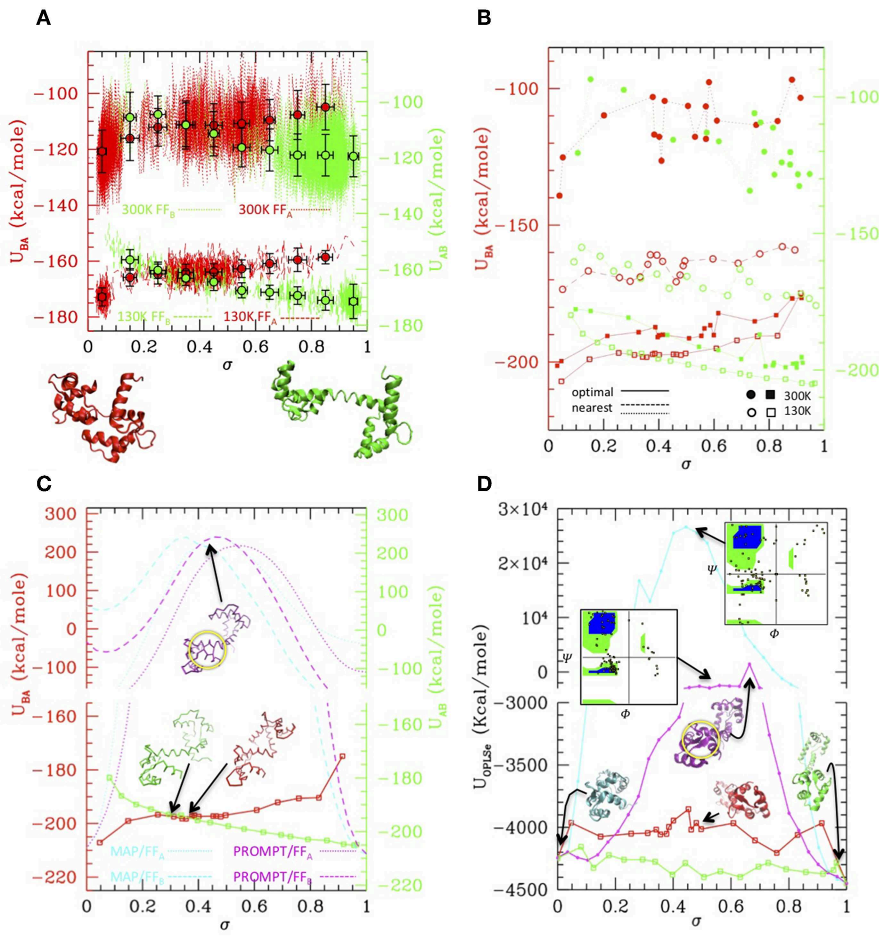


FIGURE 3 | Simulation data analysis and comparison with PROMPT and MAP **(A)** Potential energy vs. σ along the simulations at 300 K (dotted lines) and at 130 K (dashed lines), with the FF$_A$ (red) and FF$_B$ (green) force fields (scales for FF$_A$ and FF$_B$ are shifted of 3 Kcal/mole to align the activated state as explained in the text. Both scales are reported on the left and right axis, in colors corresponding to the FF they refer to). Colored dot with error bars are averages over subsets of structures classified by σ intervals (errorbars correspond to standard deviations of data from average values). Representative closed (σ = 0) and open (σ = 1) structures are reported under the plot. **(B)** Potential energies vs. σ evaluated over the representative structures of the clusters outputted by PP procedure. Squares connected by solid lines: representatives optimized by the PP procedure (filled = from the 300 K simulation, empty = from the 130 K simulations, red with FF$_A$, green with FF$_B$). Circles connected by dashed/dotted lines: same as previous, but evaluated over a trajectory of structures extracted from the simulations, the nearest to the optimal ones. (Same color and empty/filled code as for squares; shift of scales as in **A**). **(C)** Comparison of the 130 K "optimal" energies with energies of trajectories from MAP (cyan) and PROMPT (magenta) evaluated with FF$_A$ (dotted) and FF$_B$ (dashed). Representative structures of the activated states are reported in corresponding colors. Same scale shift as in **(A)**; the vertical scales are broken to zoom over the low energies. **(D)** Potential energy evaluated with the atomistic FF over the same trajectories as in **(C)** (same color coding). Representative structures are reported in corresponding colors; the Ramachandran plot of the activated states of PROMPT and MAP are reported (yellow squared dots superimposed to the standard map in colors). Both in **(C,D)** the area with distorted sheets in the activated state of PROMPT is highlighted with a yellow circle.

to provide meaningful ones, even after local optimization, in the intermediate regions. This was somehow expected: in fact stronger post-processing methods, involving e.g., the generation of swarms of unbiased trajectories from the transition states were proposed to solve this problem (Pan et al., 2008). PROMPT returns in addition good backbone local conformations along the whole path, but does not guarantee that amino-acids separated along the chain do not get too near and cause steric clashes, which happens in fact, in the intermediate regions. The MCG simulations, guarantee physically sound structures along the whole path, and can explore also intermediates far from the reference structures, but needs appropriate post-processing and clustering techniques to extract a reaction path. We envision that a synergistic use of these methods might combine accuracy and efficiency in the path search. This possibility, and the application to a number of diverse proteins, are explored in a forthcoming paper (Delfino et al., in preparation).

DATA AVAILABILITY STATEMENT

All data used for this work are included in the paper or in **Supplementary Material**.

AUTHOR CONTRIBUTIONS

FD has produced data and performed analyses, and participated in writing. YP has produced data and analyses, designed work, contributed ideas, and participated in writing. ES and GT have designed the work, contributed ideas, and participated in writing. VT has contributed ideas, designed work and supervised it, and written the paper.

FUNDING

The work of ES has been supported by the RSF grant 19-71-30020, Applications of probabilistic artificial neural generative models to development of digital twin technology for non-linear stochastic systems (HSE University). The work of YP has been supported by the Russian Academic Excellence Project 5-100 of Sechenov Medical University.

ACKNOWLEDGMENTS

We wish to thank Dr. Walter Rocchia for useful discussions and for support in using the software for PP calculations, and Prof. Paolo Carloni, Dr. Giulia Rossetti, and Dr. Emiliano Ippoliti for useful discussions on Calmodulin. We also wish to thank Vladimir Kadochnikov for help with Gromacs calculations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00104/full#supplementary-material

Numerical raw data (trajectories and energies during simulations, clusters analysis etc.) are available as **Supplementary Material**.

REFERENCES

(2019). *Schrödinger Release 2018-2: MacroModel*. New York, NY: Schrödinger LLC.

Arthur, D., and Vassilvitskii, S. (2007). "k-means++: the advantage of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms* (New Orleans, LA), 1027–1035.

Best, R. B., and Hummer, G. (2009). Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* 113, 9004–9015. doi: 10.1021/jp901540t

Boateng, H. A., and Todorov, I. T. (2015). Arbitrary order permanent Cartesian multipolar electrostatic interactions. *J. Chem. Phys.* 142:034117. doi: 10.1063/1.4905952

Bush, I. J., Todorov, I. T., and Smith, W. (2006). A DAFT DL_POLY distributed memory adaptation of the Smoothed Particle Mesh Ewald method. *Comp. Phys. Commun.* 175, 323–329. doi: 10.1016/j.cpc.2006.05.001

Di Fenza, A., Rocchia, W., and Tozzini, V. (2009). Complexes of HIV-1 integrase with HAT proteins: Multiscale models, dynamics, and hypotheses on allosteric sites of inhibition. *Proteins* 76, 946–958. doi: 10.1002/prot.22399

Evans, L. C., and Gangbo, W. (1999). Differential equations methods for the Monge-Kantorovich mass transfer problem. *Am. Math. Soc.* 653:66. doi: 10.1090/memo/0653

Ferrarotti, M. J., Rocchia, W., and Decherchi, S. (2018). Finding principal paths in data space. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 2449–2462. doi: 10.1109/TNNLS.2018.2884792

Franklin, J., Koehl, P., Doniach, S., and Delarue, M. (2007). MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acid Res.* 35, W477–W482. doi: 10.1093/nar/gkm342

Grant, B. J., Gorfe, A. A., and McCammon, J. A. (2010). Large conformational changes in proteins: signaling and other functions. *PLoS Comput. Biol.* 20, 142–147. doi: 10.1016/j.sbi.2009.12.004

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2016). OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* 12, 281–296. doi: 10.1021/acs.jctc.5b00864

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Koshevoy, A. A., Stepanov, E. O., and Porozov, Y. B. (2014). Method of prediction and optimization of conformational motion of proteins based on mass transportation principle. *Biophysics* 59, 28–34. doi: 10.1134/S0006350914010035

Masino, L., Martin, S. R., and Bayley, P. M. (2000). Ligand binding and thermodynamic stability of a multidomain protein, calmodulin. *Protein Sci.* 9, 1519–1529. doi: 10.1110/ps.9.8.1519

Pan, A. C., Sezer, D., and Roux, B. (2008). Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B.* 112, 3432–3440. doi: 10.1021/jp0777059

Polak, E., and Ribiere, G. (1969). Note sur la convergence de directions conjuguée. *Revue Francaise Inform Recherche Operat.* 16, 35–43. doi: 10.1051/m2an/196903R100351

Rabl, C. R., Martin, S. R., Neumann, E., and Bayley, P. M. (2002). Temperature jump kinetic study of the stability of apo-calmodulin. *Biophys. Chem.* 101/102, 553–64. doi: 10.1016/S0301-4622(02)00150-3

Rotkiewicz, P., and Skolnick, J. (2008). Fast method for reconstruction of full-atom protein models from reduced representations. *J. Comp. Chem.* 29, 1460–1465. doi: 10.1002/jcc.20906

Seyler, S. L., Kumar, A., Thorpe, M. F., and Beckstein, O. (2015). Path similarity analysis: a method for quantifying macromolecular pathways. *PLoS Comput. Biol.* 11:e1004568. doi: 10.1371/journal.pcbi.1004568

Spampinato, G. L. B., Maccari, G., and Tozzini, V. (2014). Minimalist model for the dynamics of helical polypeptides: a statistic-based parameterization. *J. Chem. Theory Comput.* 10, 3885–3895. doi: 10.1021/ct5004059

Tamazian, G., Chang, J. H., Knyazev, S., Stepanov, E. O., Kim, K. J., and Porozov, Y. B. (2015). Modeling conformational redox-switch modulation of human succinic semialdehyde dehydrogenase. *Proteins* 83, 2217–2229. doi: 10.1002/prot.24937

Tavanti, F., and Tozzini, V. A. (2014). Multi-scale–multi-stable model for the rhodopsin photocycle. *Molecules* 19, 14961–14978. doi: 10.3390/molecules190914961

Tirion, M. (1996). Large amplitude elastic motions in proteins from a single parameter, atomic analyses. *Phys. Rev. Lett.* 77, 1905–1908. doi: 10.1103/PhysRevLett.77.1905

Todorov, I. T., Smith, W., Trachenko, K., and Dove, M. T. (2006). DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *J. Mater. Chem.* 16, 1911–1918. doi: 10.1039/B517931A

Tozzini, V. (2005). Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15, 144–150. doi: 10.1016/j.sbi.2005.02.005

Tozzini, V. (2010). Minimalist models for proteins: a comparative analysis. *Q. Rev. Biophysis.* 43, 333–371. doi: 10.1017/S0033583510000132

Tozzini, V., and McCammon, J. A. (2005). A coarse grained model for the dynamics of flap opening in HIV-1 protease. *Chem. Phys. Lett.* 413, 123–128. doi: 10.1016/j.cplett.2005.07.075

Tozzini, V., Rocchia, W., and McCammon, J., A. (2006). Mapping all-atom models onto one-bead coarse-grained models: general properties and applications to a minimal polypeptide model. *J. Chem. Theory Comput*. 2, 667–673. doi: 10.1021/ct050294k

Tozzini, V., Trylska, J., Chang, C., and McCammon, J. A. (2007). Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *J. Struct. Biol.* 157, 606–615. doi: 10.1016/j.jsb.2006.08.005

Trovato, F., Nifosì, R., Di Fenza, A., and Tozzini, V. (2013). A minimalist model of protein diffusion and interactions: the green fluorescent protein within the cytoplasm. *Macromolecules* 46, 8311–8322. doi: 10.1021/ma401843h

Trovato, F., and Tozzini, V. (2012). Minimalist models for biopolymers: open problems, latest advances and perspectives. *AIP Conf. Proc.* 1456, 187–200. doi: 10.1063/1.4730659

Weis, W. I., and Kobilka, B. K. (2008). Structural insights into G-protein-coupled receptor activation. *Curr. Opin. Struct. Biol.* 18, 734–740. doi: 10.1016/j.sbi.2008.09.010

Weiss, D. R., and Levitt, M. (2009). Can morphing methods predict intermediate structures? *J. Mol. Biol.* 385, 665–674. doi: 10.1016/j.jmb.2008.10.064

Wenfei, L., Wang, W., and Takada, S. (2014). Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10550–10555. doi: 10.1073/pnas.1402768111

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.