Exploration of conformational transition pathways from coarse-grained simulations

Pedro Sfriso\textsuperscript{1,2}, Adam Hospital\textsuperscript{1,2,3}, Agustí Emperador\textsuperscript{1,2} and Modesto Orozco\textsuperscript{1,2,3,4,*}

\textsuperscript{1}Institute for Research in Biomedicine (IRB Barcelona), \textsuperscript{2}Joint IRB-BSC Program in Computational Biology, Baldiri Reixac 10, Barcelona 08028, Spain, \textsuperscript{3}National Institute of Bioinformatics and \textsuperscript{4}Department of Biochemistry and Molecular Biology, University of Barcelona, Av. Diagonal 647, Barcelona 08028, Spain

\textbf{ABSTRACT}

\textbf{Motivation:} A new algorithm to trace conformational transitions in proteins is presented. The method uses discrete molecular dynamics as engine to sample protein conformational space. A multiple minima Go-like potential energy function is used in combination with several enhancing sampling strategies, such as metadynamics, Maxwell Demon molecular dynamics and essential dynamics. The method, which shows an unprecedented computational efficiency, is able to trace a wide range of known experimental transitions. Contrary to simpler methods our strategy does not introduce distortions in the chemical structure of the protein and is able to reproduce well complex non-linear conformational transitions. The method, called GOdMD, can easily introduce additional restraints to the transition (presence of ligand, known intermediate, known maintained contacts,…) and is freely distributed to the community through the Spanish National Bioinformatics Institute (http://mmb.irbbarcelona.org/GOdMD).

\textbf{Availability:} Freely available on the web at http://mmb.irbbarcelona.org/GOdMD.

\textbf{Contact:} modesto.orozco@irbbarcelona.org or modesto@mmb.pcb.ub.es

\textbf{Supplementary information:} Supplementary data are available at Bioinformatics online.

Received on February 26, 2013; revised on April 29, 2013; accepted on May 30, 2013

\section{INTRODUCTION}

Many biological functions of proteins such as mechanic work, signal transduction or enzymatic activity are modulated by a key property of them: flexibility (Henzler-Wildman \textit{et al.}, 2007; Micheletti, 2013; Velazquez-Muriel \textit{et al.}, 2009). Flexibility is a property that has been refined and maintained by evolution (Falke, 2002; Micheletti, 2013) and that, in turn has been also exploited by evolution to generate new proteins in a conservative mechanism, which guarantees the maintenance of the structural scaffold as well as the relevant deformation pattern (Leo-Macias \textit{et al.}, 2005; Stein \textit{et al.}, 2011). Structural databases show (Gerstein and Krebs, 1998) increasing number of proteins having alternative structures depending on external factors (such as crystallization conditions, posttransductional chemical modifications, presence of ligands, changes in solvent environment, etc). This probes the existence of dramatic conformational transitions in proteins, but giving no information on how such transitions happen.

Recent refinements of experimental techniques have provided direct evidence on the mechanisms of conformational transitions for some model proteins (Ban \textit{et al.}, 2011; Eisenmesser \textit{et al.}, 2002; Fenwick \textit{et al.}, 2011; Kern and Zuiderweg, 2003; Lindorff-Larsen \textit{et al.}, 2005). However, we are still far from the point that all conformational transitions could be described by means of experimental methods. This situation forces the use of simulation techniques, which has been largely refined in the past years (Best \textit{et al.}, 2005; Bolhuis \textit{et al.}, 2002; Kaprlish and Kuriyan, 2005; Kubitzki and de Groot, 2008; Maragakis and Kaprlish, 2005; Miyashita, 2003; Okazaki \textit{et al.}, 2006), providing information of increasing quality for a large number of conformational transitions in proteins (Sfriso \textit{et al.}, 2012; Stein \textit{et al.}, 2011).

Molecular dynamics (MD), using atomistic force-fields and explicit representation of solvent (McCammon \textit{et al.}, 1977), is probably the most accurate theoretical technique for reproducing protein flexibility. Recent computational approaches have made possible the simulation of thousands of proteins (Meyer \textit{et al.}, 2010) and the derivation of up to millisecond trajectories for proteins (Dor \textit{et al.}, 2012; Lindorff-Larsen \textit{et al.}, 2011). Unfortunately, most conformational transitions are still far from the capabilities of plain atomistic MD, forcing the use of biasing schemes (Beckstein \textit{et al.}, 2009; Das \textit{et al.}, 2006; Laio and Parrinello, 2002; Leone \textit{et al.}, 2010; Liphardt \textit{et al.}, 2002; Schlitter, 1994), designed to maximize the sampling along a given variable that is believed to capture conformational transition motions. Biased-MD protocols are extremely powerful, but they are expensive computationally, require expertise from the user and can lead to incorrect results when the transition coordinate is not well defined.

Coarse-grained (CG) models (Bahar and Rader, 2005; Dobbins \textit{et al.}, 2008; Lopez-Blanco \textit{et al.}, 2011; Mendez and Bastolla, 2010; Orozco \textit{et al.}, 2011; Tocchini, 2005; Whitford \textit{et al.}, 2007) are inexpensive, but still accurate alternatives to atomistic MD simulations, which have gained significant popularity in recent years. Unfortunately, the use of CG models require the assumption of a certain loss of detail in the simulation, for example, explicit solvent is ignored, which prevent the study of specific water-protein interactions and side chains are either ignored or dramatically simplified (Mendez and Bastolla, 2010; Marrink \textit{et al.}, 2007), which raises problems.

*To whom correspondence should be addressed.
to reproduce ligand-target interactions. CG models have been calibrated against experimental data, such as B-factors, structural variability in databases or atomistic MD, and despite their simplicity, provide often results of surprising quality with limited computer resources and reduced expertise from the user.

CG-methods have been largely used to trace conformational transitions in proteins. The first implementations followed mainly interpolation schemes between original and final conformations (Delarue and Sanejouand, 2002; Flores et al., 2006; Franklin et al., 2007; Kim et al., 2002; Krebs and Gerstein, 2000; Lindahl et al., 2006; Weiss and Levitt, 2009). Interpolation protocols are fast and guarantee the completion of the transition, but often at the expense of unrealistic intermediate conformations, which are not good starting points for refinement through more accurate atomistic simulations. Alternative CG-transition methods have been developed under the assumption that biologically relevant transitions should follow the easiest deformation movements of the proteins (Bahar et al., 2010; Lezon et al., 2009; Mendez and Bastolla, 2010; Yang et al., 2009) defined as the softest deformation modes obtained by diagonalization of the Hessian matrix derived from an elastic network model (ENM) Hamiltonian:

$$E = \sum_{i,j} \delta_{ij} K_{ij} \left( R_{ij} - R_{ij}^0 \right)^2$$

where \( i \) and \( j \) are residues, \( \delta_{ij} \) is a delta function equal to 1 when \( i \) and \( j \) are at less than a given distance, and 0 otherwise; \( K \) is a spring constant (linear or distance dependent); \( R_{ij} \) stands for inter-residue distance and the superscript 0 refers to the value of \( K_{ij} \) in the reference structure.

Morphing methods based on the use of essential deformation modes provide more reasonable approaches than linear interpolation schemes, but are far from being optimal since (i) the covariant structure of the protein can be damaged when large displacements along a limited number of Cartesian eigenvectors are made, (ii) a non-negligible number of transitions do not align with the intrinsic deformation pattern of proteins (Stein et al., 2011), and even those cases where the alignment between the essential deformation space and the transition vector is good, displacement along the essential space of one of the proteins never allows a full transition between the two conformational states. Recent advances in the field, such as the use of multiple reference structures to define a transition-dependent essential deformation space (Sfriso et al., 2012; Yang et al., 2009) or the derivation of the essential deformation modes in the dihedral space (Lopez-Blanco et al., 2011; Mendez and Bastolla, 2010), have alleviated but not solved these problems.

In this article, we present a new approach to obtain ultra-fast, but accurate, conformational transition pathways in proteins out of physics-based molecular mechanics simulations. By combining an efficient sampling technique [discrete molecular dynamics (dMD; Orozco et al., 2011; Proctor et al., 2011)] with a novel multiple-well Go-like scheme. Sampling is enhanced with biasing techniques such as metadynamics (Laio and Parrinello, 2002) and Maxwell Demon MD (Rueda et al., 2004; Sfriso et al., 2012). The method was tested in a large battery of transitions (near 50 pairs of structures) obtaining, with reduced computational cost, reasonable pathways in all cases, including in cases of extreme difficulty, where transition requires, for example, partial unfolding of the protein. Owing to the physical nature of the method, sampled intermediate structures maintain covalent structure, and no steric clashes are allowed. In cases where experimental intermediates are characterized, the method finds transition paths passing close to them. Surprisingly, the method outperforms our previous more detailed algorithm, Maxwell Demon discrete molecular dynamics procedure (MDdMD; Sfriso et al., 2012), and is also competitive with respect to alternative methods in the literature (Supplementary Figs S1–S3). Our novel approach is flexible, allowing the introduction of any perturbation. Furthermore, the user can easily bias trajectories to guarantee that known experimental intermediates are sampled.

2 METHODS

In this approach we have applied the dMD method [see (Emperador et al., 2008b; Proctor et al., 2011) for details]. In dMD, particles move in the ballistic regime, with constant velocity until a collision occurs. Collisions occur at the particle–particle distances corresponding to a discontinuity in the dMD potential. The velocities of the particles after the collision are obtained using the rules of conservation of momentum and energy. This allows avoiding the integration of Newton’s equations of motion, speeding up the simulations as compared with usual MD. Discrete MD has successfully been applied to protein folding (Ding et al., 2005; Zhou and Karplus, 1997), macromolecular dynamics (Emperador et al., 2008a, 2010), RNA structural predictions (Ding et al., 2008, 2012; Gherghethe et al., 2009) and protein aggregation (Ding and Dokholyan, 2008; Urbanc et al., 2004, 2010). Recently dMD has been also used to robust energy minimization of protein–protein complexes (Emperador et al., 2013) and protein–ligand interactions (Proctor et al., 2012).

2.1 Force-field representation

We adopted a CG representation of the protein, where only Cαs were explicitly represented. We used a single square well with infinite walls to define the bonded interactions between consecutive beads. The non-bonded interactions were described by a multiple well Go-like model (Taketomi et al., 1975; Ueda et al., 1978) depending on the experimental inter-particle distance. Non-bonded terms of the Hamiltonian were classified in two categories based on whether or not the inter-particle distance is different between the initial (A) and final (B) conformations. If they are, we define a double well potential (Fig. 1) centered in the respective experimental values: \( r_A \) and \( r_B \). When the distances are similar in both conformations (i.e. \( r_A \sim r_B \)), we used a single, but wider, well (50% larger) centered at \( r_A + r_B \). To avoid over-restraining the system with physically irrelevant interactions, the potential energy was defined only for those particles within a cutoff \( (r_A + r_B) > \text{radius of gyration}/2 \). It is worth noting that the height of the wells for all non-bonded interactions is finite (0.5 kcal/mol), allowing the particles to escape from the wells if required.

This Go-like scheme is convenient to complete transition paths (>85% of the root mean square deviation (RMSD) difference between end structures) because target minimum acts as attractor. This situation enables us to use softer biasing schemes and to better recover native contacts (see Supplementary Materials for details).

2.2 Accelerating the transition

Within the Hamiltonian definition above, the A–B transition occurs through a movement of particles jumping from wells centered at \( r_A \) to those centered at \( r_B \). Unfortunately, as a side effect of Go-like potentials,
Two known structures (EGO 1982, P.Sfriso et al. 2012) were signed to favor the transition, but avoiding the use of restraints that would guide the trajectory along arbitrary (and probably unrealistic paths). The first level of biasing is designed to move the protein far from the starting conformation, while the second level is designed to enrich biasing to approach the protein to the final conformation. For the interested reader, a flowchart of the algorithm is presented in Supplementary Figure S4.

2.2.1 Escaping from the initial minima We have implemented a discrete version of the metadynamics method (Laio and Parrinello, 2002) to guarantee that the system leaves the original minima. Metadynamics penalizes visits to the original well by raising gradually the energy of the well (i.e. by filling it; Fig. 1), which increases the chances of the system to escape from the attraction of the Go-potentials to the starting structure. As a consequence, as trajectory progresses sampling of the original conformation is less and less probable and system departs from the original conformation. Metadynamics is efficient as a method to escape from a minimum, but such a divergence happens in a random way, which can yield to unrealistic deformations when applied in an unsupervised fashion. To solve this problem, we have coupled metadynamics to an ENM (Emperador et al., 2006a; Orellana et al., 2010) in such a way that wells associated to inter-particle distances showing large changes along the first five essential deformation modes were filled faster than those associated to inter-particle distances that are not coupled to the essential deformation modes of the protein (Fig. 1). This strategy guarantees the exploration of alternative ways to escape from the original structure minima, while increasing the possibility to sample preferred pathways as defined by the essential deformation modes (first five are considered by default, but results are largely invariant in the range 3–10 modes; data not shown). Note that the ENM analysis is done only with the initial structure irrespectively of which the target structure is, avoiding the definition of too linear pathways. Note also that the full dMD simulations are done, which means that all the degrees of freedom (and not only the five preferred ones) are sampled.

2.2.2 Moving toward the final conformation The ENM-metadynamics procedure outlined above guarantees that the protein moves apart from the original structure sampling preferably essential deformation modes. However, there is no guarantee that such movements will approach the protein to the final conformation. To guide the trajectory toward the final structure, without introducing arbitrary energy restraints, we have implemented a Maxwell Demon biasing algorithm (Perilla et al., 2010; Rueda et al., 2004; Sfriso et al., 2012) with a control magnitude (Γ) defined as follows:

$$\Gamma = \sum_{i=1}^{N} \omega(i) \left| r_{i,B} - r_{i,X} \right|$$  \hspace{1cm} (2)

where N is the total number of residues, B is the target structure, X is the sampled conformation (X = A for the original conformation) and ω(i) is an optimized weight function dependent on the inter-particle distance and the size of protein (Supplementary Fig. S5; results are robust to ω values in the range 15–25 Å are used).

Following the MDDMD procedure the bias toward the target structure is not introduced by an energy penalty, but using a less interfering informational criteria (Perilla et al., 2010). The scheme is simple and efficient, after a certain number of dMD simulation steps (time t) the value of the progress variable (Γt; equation 2) is compared with that obtained in a previous accepted movement (t−Δt). The simulation fragment t−Δt→t is then accepted or rejected based on a simple Metropolis test (equation 3):

$$p_t = \begin{cases} 1 & \text{if } \Gamma_t < \Gamma_{t-\Delta t} \\ \exp\left(-\frac{(\Gamma_t - \Gamma_{t-\Delta t})^2}{2\beta}\right) & \text{if } \Gamma_t > \Gamma_{t-\Delta t} \end{cases}$$  \hspace{1cm} (3)

where p is the acceptance probability, RMSd_{A,X,t} is RMSd between structure sampled at time t(X) and target structure, β is dynamically

Fig. 1. Interaction potentials used in this study. (a) single well corresponding to a particle–particle non-bonded pair that does not change their distance during transition (E_{GO} = 0.5 kcal/mol, σ_g = 0.15 Å). (b) Double well representing two states of particle radial distance that corresponds to two known structures (E_{GO} = 0.5 kcal/mol, σ_g = 0.10 Å). (c) Example of how discrete metadynamics increases the potential energy of a pair of particles whose movement overlaps with the essential deformation space of the initial state and (d) the same situation but when the overlap is much lower.
adjusted to guarantee an acceptance rate of 70%, and the time frame (Δt) is typically 100 time steps. These large acceptance rates, combined with a weight function with a maximum at ~15 Å enables the system to explore no so direct pathways that could lead to stressed structures and provides the best local geometries (Supplementary Figs S6 and S7). The presence of the Go-potentials of the target structure in the Hamiltonian avoids the need to increase β when trajectory approaches the final target conformation.

3 RESULTS

To evaluate the power of GOdMD, we apply it to trace transitions between known equilibrium conformations of proteins. After analyzing protein data bank (PDB), we define a set of 94 transitions corresponding to 47 proteins showing two distinct conformations. This large benchmark set was extracted from a previous work in our group (Sfriso et al., 2012) and spans a wide range of proteins from 100 to 1000 residues, showing different shapes and secondary structure composition. The database contains no trivial conformational changes and some of the conformational transitions represent dramatic geometrical alterations in the structure of the protein, including refolding in some cases. Many of the transitions are coupled to the binding to small ligands (60%) and/or macromolecules (39%), which increase the difficulties to trace reasonable conformational transition pathways. Further analysis of the data set reveals that a significant fraction of the transitions (44%) correspond to open/close ones, which can a priori generate hysteresis problems. Finally, 22% of the transitions do not align well [overlap (accumulated dot product) between top five normal modes and conformational transition is below 0.25] with the first five essential deformation modes detected from EN-normal mode analysis (NMA), i.e. that they will be very difficult to represent by simply activating movements along intrinsic deformation modes. Looking at different criteria (Sfriso et al., 2012) we consider 62 of the 94 transitions as difficult or very difficult to follow, which means that we are validating our method with the most exigent transition dataset available.

GOdMD has been able to find reasonable pathways for all studied transitions, even those requiring large refolding processes. The algorithm can provide multiple transition trajectories, defining a scenario of multiple pathways, which fits better into the transition funnel theory (Dill and Chan, 1997; Portella and Orozco, 2010). Furthermore, no violations of covalent distances, nor steric clashes or chemical meaningless conformations are sampled during transitions, which are always smooth, without the presence of apparent discontinuities or hysteresis effects (Supplementary Fig. S8). PROSA (Wiederstein and Sippl, 2007) calculations performed in random conformations sampled during the different transition yields to native-like profiles (Supplementary Figs S9 and S10), indicating that we are not sampling unrealistic conformations along the transition. The method is extremely efficient, >60% of trajectories are finished in <2 min wall-clock time in a laptop computer (2.4 GHz Intel Core 2 Duo) and even the most difficult transitions are finished in <30 min in the same computer. We are working in a parallel algorithm to explore larger systems with similar efficiency.

3.1 Sampling known intermediate conformations

There are a few cases in PDB of distinct conformations for a protein, where there is in addition to the start and end conformation a third structure, which is intermediate (at least in terms of RMSd) between the other two. Following Weiss and Levitt (2009) we can suggest that in general the third structure can be near a preferred passing point on the transition between the extreme conformations (i.e. we can consider an ‘intermediate’ in the transition). Thus, for the five cases where this putative intermediate is available, we determined transition paths between the two more diverse conformations. In all cases (Fig. 2) we found smooth and reasonable transitions, which go in the direction of the putative intermediate, even for the most difficult cases, where simpler methods have serious difficulties. In small systems (Ribose Binding Protein, 5'-NT), the transition passes through the putative intermediate. In larger systems, small RMSd deviations to the putative intermediate are much more difficult to obtain, partly because of noise introduced by the presence of
long connecting loops, and partly because the intrinsic difficulty of tracing large conformational transitions. Although some large RMSd values are obtained (due mostly to stochastic motion movements), the potentials used in GOhMD capture better intermediate states than other state-of-the-art morphing methods (Supplementary Fig. S3). However, it is clear that the simulation drives spontaneously the transition toward conformational regions close to the putative intermediate. PROSA profiles and TMcore (Zhang and Skolnick, 2004) of both sampled and putative experimental intermediates are similar (Supplementary Figs S9 and S10 and Supplementary Table S2 for details), supporting the quality of the transition path.

3.2 Non-linearity of the transitions

One of the main caveats of biasing methods based on forcing a regular reduction of RMSd to the target structure (as most morphing methods) is that they force a linear transition path, that might be unrealistic (Fig. 3), and that does not provide information on the bottlenecks in the transition. Our method is able to detect non-linear transition pathways, and provides information on where are the bottlenecks, i.e. those points where the transition seems nearly stopped for long periods, and which correspond to regions of high rejection rate in the Maxwell Demon (Supplementary Fig. S11). Interestingly, the method is also able to disconnect local (measured as % of native contacts) to global transitions (measured by the RMSd to the target), showing how for some transitions local rearrangements happen first, while, on the contrary, for other transitions global movement is before local conformational rearrangements (Figs 3 and 4).

It is worth noting that, the physical nature of the method guarantees that the guiding engine (NMA-bias metadynamics and a Maxwell Demon here) finds conformational pathways that never explore regions of unrealistically large energy. This guarantees that along all the transitions we sample protein-like conformations (Fig. 4), avoiding sampling always linear pathways that can leads to strong local distortions.

3.3 Introducing experimental information

One of the main advantages of our method is its flexibility, which makes very easy to introduce the effect of external perturbations in the transition, or to bias the trajectory by experimental information.

To show the capabilities of our algorithm, we analysed the effect of ligand binding in a transition, and also illustrate how the existence of experimentally validated intermediate can be used to bias the transition towards preferred pathways. For the first purpose we studied three additional systems for which ligand binding is known to displace conformational equilibrium: D-Allose binding protein (1gud/1rpj), L-Leucine Binding Protein (1usg/1usi) and Osmo-protection protein (1sw5/1sw2). After running 10 simulations (five in the absence and five in the presence of the ligand), we observed that those trajectories containing ligand were less efficient to reach the final state (Supplementary Figs S12 and S13 for details). These are just ‘prove of concept’ calculations, but provide encouraging evidence that our simple and fast algorithm can capture qualitatively the effect of the ligand in altering conformational pathways. To explore the possibility to bias GOhMD transitions by using experimental information on intermediate we have repeated the study of the very difficult 1SU4 → 1WO transition (Weiss and Levitt, 2009) assuming that the intermediate structure 1VFP corresponds to a necessary pass-point in the transition. So, we decoupled our sampling methods to find the transition pathways that better accommodate this experimental information.

With the same Hamiltonian defined by the end states (A, B; see Methodological Approach), the Maxwell Demon MD is now first referred to the experimental intermediate structure (1VFP) and then switched to the target structure (when sampling of intermediate structure is converged). With this protocol, we selected the path closer to the putative intermediate state I. Figure 5 shows the obtained results. A clear improvement on the sampling of I is observed, showing that while the potential

Fig. 3. (a) Energy surface where a linear conformational transition is expected to be realistic (x and y correspond to two general conformational variables, and color code means the energy order: red-yellow-green-blue). (b) Energy surface where transition does not follow a linear pathway. (c) An example of a real transition (4ake→1ake) that behaves linearly, and where pathway obtained by our current procedure (black line) re a linear one (blue; obtained by forcing low acceptance rates and ignoring ENM analysis). (d) As in panel (c), but for a non-linear transition 1c9kB→1cbuB, where normal (black) and forced linear transitions (blue) are different, the latter being incorrect, because little target-native contacts are recovered (despite the low RMSd to target).

Fig. 4. Typical RMSd to target and potential energy profiles. Energy is presented in relative units. Different (three for simplicity) trajectories lead to similar but not identical results.
implementing our method is publicly available as a web server at mmb.irbbarcelona.org/GOdMD.

**Funding** Ministry of Economy and Competitiveness, Spain, the European Research Council (ERC-Advanced Grant (BIO2012-32868) to M.O.); the Instituto Nacional de Bioinformatica (INB); the Consolider E-Science Project (to M.O.); Framework VII Scalalife Project (to M.O.); Fundación Marcelino Botín (to M.O.). P.S is a ‘La Caixa fellow’ and M.O. is an ICREA Academia Researcher.

**Conflict of Interest** none declared.

**REFERENCES**


Ding, F. et al. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA, 14, 1164–1173.


Fig. 5. 1SU4 \rightarrow 1IWO transition including (blue) or not (black) information on the existence of an intermediate 1VFP Transition is repeated five times to guarantee reproducibility. (a) Profiles of RMSd with respect to experimental intermediate structure. (b) Potential energy evolution along the trajectory.


