Bioinformatics Advance Access published December 16, 2015

CG2AA: Backmapping Protein Coarse-Grained Structures

Leandro E. Lombardi [1], Marcelo A. Martí [2] and Luciana Capece[3,*]

[1]Dto. de Matemática - Instituto de Cálculo, [2]Dto. de Química Biológica - IQUIBICEN,[3]Dto. de Química Inorgánica, Analítica y Química Física - INQUIMAE,[1,2,3] Fac. de Ciencias Exactas y Naturales, Univ. de Buenos Aires, Cdad. Universitaria, C1428EGA, CABA, Argentina.

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Summary: Coarse-grained (CG) models allow long-scale simulations with a much lower computational cost than that of all-atom (AA) simulations. However, the absence of atomistic detail impedes the analysis of specific atomic interactions that are determinant in most interesting biomolecular processes. In order to study these phenomena, it is necessary to reconstruct the atomistic structure from the CG representation. This structure can be analyzed by itself or be used as an onset for atomistic molecular dynamics simulations. In this work we present a computer program that accurately reconstructs the atomistic structure from CG models from the force field developed in (Hills *et al.*, 2010), using a simple geometrical algorithm.

Availability: The software is free and available online at http://www.ic.fcen.uba.ar/cg2aa/cg2aa.py

Contact: lula@qi.fcen.uba.ar

Supplementary information: More detail on the algorithm is available as online-only supplementary information at the journal's web site. Also, a tutorial and an example can be found at http://www.ic.fcen.uba.ar/cg2aa/cg2aa.html

1 INTRODUCTION

In the context of molecular dynamics, coarse grain (CG) models define particles by grouping several atoms into one *bead*. This process reduces the number of particles used to describe the system, allowing longer scale (either temporal, spatial or both) simulations in exchange of detail (Saunders and Voth, 2013). Furthermore, the possibility of reconstructing (or *backmapping*) the atomistic structure from the CG representation allows a multiscale approach. Although a perfect reconstruction is mathematically impossible, approximate reconstructions can be obtained (Heath *et al.*, 2007; Rzepiela *et al.*, 2010; Wassenaar *et al.*, 2014; Darré *et al.*, 2015). For instance, in (Wassenaar *et al.*, 2014), authors construct a complete backmap for the Martini model (Monticelli *et al.*, 2008). This reconstruction is remarkably accurate for the protein backbone and quite good on the side chains.

The backmapping algorithm presented here was developed for the CG model developed in (Hills *et al.*, 2010). As in the Martini force field, this model represents the backbone with one bead. Nevertheless, this bead is located in the position of the C_α atom instead of the center of mass of the residues backbone atoms. The side chains are modeled with 1 to 4 beads, depending on the complexity of the aminoacid. The backbone reconstruction scheme presented in this work can be applied to any model that represents the backbone with a bead at the C_α. Of course, side chains reconstructions strongly depend on number and positions of the beads of each CG model but the same approach can be implemented in other models (see Fig. S4 for a Martini CG model example). We also include in our program a backmapping of a CG model for the heme group, which was obtained using a similar approach as the one used to obtain standard CG aminoacids in (Hills *et al.*, 2010).

In summary, this tool allows backmapping of CG structures in an accurate manner. Also, due to the simplicity of the algorithm, it can be easily expanded to other particular molecules or chemical groups, such as cofactors and ligands (e.g. the heme group).

2 METHODS

The algorithm is implemented in Python, foreseeking maximum portability. Numerical calculations are performed with the NumPy library, a nowadays standard in any python-enabled system. The software usage is very simple by invoking the program from the command line.

The algorithm works as follows. First, we reconstruct backbone atoms positions from the known C_α positions. In (Wassenaar *et al.*, 2014) the authors claim that the C=O vector of the i-th amino acid points approximately in the direction of the cross product of the vectors $\overrightarrow{C_\alpha^i C_\alpha^{i+2}}$ and $\overrightarrow{C_\alpha^i C_\alpha^{i+1}}$. We propose a weaker assumption: This cross product lies inside the peptide bond plane (see Fig. 1a). Since this plane contains both C_α^i and C_α^{i+1}, the positions of the atoms between them are completely determined. More precisely, we place the C_i, O_i and N_{i+1} atoms at distances given by the Amber force field (Duan *et al.*, 2003) and the angles reported in (Tozzini *et al.*, 2006) (see Supplementary Information (SI) for details). This gives an accurate reconstruction, except at the N- and C-terminals.

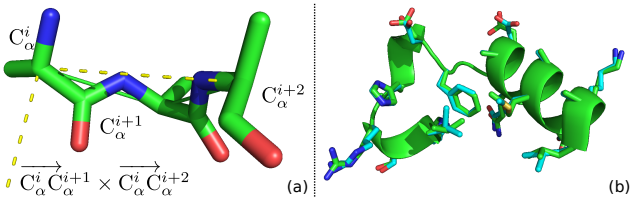

Fig. 1. (a) Three consecutive C_α's and the corresponding cross product. This vector and the C=O bond have different directions. (b) In this example, the backmapped structure (cyan) is compared the original (green).

Having placed the backbone atoms, the position of the C_β (if present) is defined by the tetrahedral arrangement of the atoms bonding the C_α. Now, knowing the coordinates of the C_α atoms and, having stated the C_β positions, we guess the remaining positions of the side chain sequentially. The program computes them as follows. The already determined atoms impose geometric constrains on the positions of the ones yet to place. As expected, the guess is strongly determined by the position of the bead in question. For example, in this model most aminoacids have a bead defined by the C_β, the C_γ and the neighboring hydrogen atoms. In that case, the first step of the side chain reconstruction is to place the C_γ, given the positions of the C_α and C_β. Since the vector $\overrightarrow{C_\beta C_\gamma}$ points approximately in the direction of the bead, we chose to place the atom at a typical distance for this bond, in a way that it forms a tetrahedral angle and pointing as close as possible in that direction. The process continues atom by atom, taking into account the geometrical information of the aminoacid beads For instance, aromatic rings planes can be reconstructed from bead information. For more details on the backmapping algorithm see SI, particularly, Figs. S1 and S2.

By similar geometric reasoning, the algorithm reconstructs the atom positions from the CG model of the heme group. This model was obtained using an analog approach as the one used to obtain standard CG aminoacids in (Hills *et al.*, 2010). In this particular case, it involves 17 beads, 1 bead for the Fe atom, 12 on the heme plane and 4 beads for the propionate side chains. (Fig. S3) The amount of beads used for its description, together with its simple geometry allows us to give an accurate reconstruction.

The numerical values of all the geometric constraints (i.e. bond distances, angles and dihedrals values) were taken from the Amber03 Force Field, (Duan *et al.*, 2003).

In order to refine the internal coordinates of the side chains and since our motivation is to use the all-atom structure as a starting point of a molecular dynamics simulation, we include an energy minimization as part of our standard work flow. Namely, we run a short classical minimization using the Amber14 package (Case *et al.*, 2015) in implicit solvent. (See SI for simulation details)

The minimization process is not included as part of this piece of code since it would severely affect portability and, at the same time, users may prefer a different simulation context (i.e. force field or software). We are confident that a user with experience in molecular dynamics simulations will find our contribution easy and flexible to use and adapt.

As examples, we have selected five different proteins with different secondary structure elements and variable three-dimensional folds. In all cases, a CG representation is obtained from the original PDB structures using an adapted version of the script gently provided by the authors of (Hills *et al.*, 2010) (see SI for details). Then, we applied the backmapping algorithm to the CG structure. The backmapped structure was compared to the original crystal structure. The total root mean square deviation (RMSD) of the backbone atoms between the backmapped and the crystal structure ranged from 0.5 Å to 0.8 Å in the analyzed cases. These values were not significantly modified after minimization. However, in some cases, minimization allowed to correct small imperfections in the relative position of the side chains with respect to their direct neighbors. RMSD calculations for each residue show very small differences for most of the protein structure. Only a few residues located, as expected, in regions with low secondary structure organization had higher deviations (see Fig. S4). Two of the selected examples correspond to heme proteins, in which the performance of the heme group backmap was also tested (the reconstruction being accurate on the heme plane, with a total RMSD of 0.61 Å).

In order to test the stability of the obtained structures, we used them as starting points for molecular dynamics (MD) simulations. More precisely, after solvation and a short thermalization, a 100-ns molecular dynamics simulation was performed for two of the considered examples. In all cases structures remained stable showing an average RMSD with respect to the original crystal structure of 1 to 2.5 Å for the backbone atoms, and lower than 3 Å when considering all the protein atoms (Fig. S6). The obtained trajectories were compared with MD simulations starting from the crystal structures, calculating the root-mean-square-fluctuation (RMSF) for each residue in the protein, showing similar results. (Fig. S7) To further test the backmapping scheme, we backmapped 500 ns a CG-MD trajectory and calculated the RMSD with respect to the crystal structure. The average RMSD resulted 3.289 Å for the backbone atoms and 4.085 Å for all the heavy atoms, which results a reasonable value considering the increased sampling and the expected deviations from the crystal structure in a CG-MD simulation (Hills *et al.*, 2010). From this backmapped CG-MD trajectory, 10 equally-spaced snapshots were simulated for 10 ns in explicit solvent. Interestingly, the RMSF obtained in these trajectories is in good agreement with the fluctuations observed in atomistic simulations starting from the corresponding crystal structure, with, again, higher fluctuations. These results further validate the application of this backmapping methodology for combining CG and atomistic simulations in a multiscale approach. (Fig. S8)

3 CONCLUSION

We presented an intuitive, portable, fast, free and easy-to-use method to map CG structures back to AA representation. We have tested it in several examples by changing protein structures to CG representation and backmapping them. Examples were selected in order to test the algorithm for different secondary structure elements and 3D structure complexity. Both the backbone and total reconstruction is accurate except, sometimes, at the more flexible parts of the structure (i.e. ends, loops and long side chains).

ACKNOWLEDGEMENT

The authors want to thank A. Roitberg for valuable discussions. They would also like to acknowledge M. Arrar and L. Boechi for helpful comments on the manuscript.

Funding: The authors are members of CONICET. This work has been supported by grants UBA-CYT project nbr. 20020120300025BA and ANPCYT-PICT 2012-2571.

REFERENCES

Case, D. *et al.* (2015). Amber 15. *University of California, San Francisco*.

Darré, L. *et al.* (2015). Sirah: A structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J. Chem. Theory Comput.*, **11**, 723–739.

Duan, Y. *et al.* (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**, 1999–2012.

Heath, A. *et al.* (2007). From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins: Struc., Funct. Bioinf.*, **68**, 646–661.

Hills, R. D. *et al.* (2010). Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.*, **6**, e1000827.

Monticelli, L. *et al.* (2008). The martini coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.*, **4**, 819–834.

Rzepiela, A. *et al.* (2010). Reconstruction of atomistic details from coarse-grained structures. *J. Comput. Chem.*, **31**, 1333–1343.

Saunders, M. and Voth, G. (2013). Coarse-graining methods for computational biology. *Annu. Rev. Biophys.*, **42**, 73–93.

Tozzini, V. *et al.* (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.

Wassenaar, T. A. *et al.* (2014). Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.*, **10**, 676–690.