Published online 4 June 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W657–W661***
doi:10.1093/nar/gkq498

SMOG@ctbp: simplified deployment of structure-based models in GROMACS

Jeffrey K. Noel[1], Paul C. Whitford[2], Karissa Y. Sanbonmatsu[2] and José N. Onuchic[1,*]

[1]Center for Theoretical Biological Physics and Department of Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093 and [2]Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, MS K710, Los Alamos, NM 87545, USA

Received February 12, 2010; Revised May 14, 2010; Accepted May 18, 2010

ABSTRACT

Molecular dynamics simulations with coarse-grained and/or simplified Hamiltonians are an effective means of capturing the functionally important long-time and large-length scale motions of proteins and RNAs. Structure-based Hamiltonians, simplified models developed from the energy landscape theory of protein folding, have become a standard tool for investigating biomolecular dynamics. SMOG@ctbp is an effort to simplify the use of structure-based models. The purpose of the web server is two fold. First, the web tool simplifies the process of implementing a well-characterized structure-based model on a state-of-the-art, open source, molecular dynamics package, GROMACS. Second, the tutorial-like format helps speed the learning curve of those unfamiliar with molecular dynamics. A web tool user is able to upload any multi-chain biomolecular system consisting of standard RNA, DNA and amino acids in PDB format and receive as output all files necessary to implement the model in GROMACS. Both C_α and all-atom versions of the model are available. SMOG@ctbp resides at http://smog.ucsd.edu.

INTRODUCTION

It is well established that the dynamic properties of biomolecules are important for their biological function. Conformational rearrangements are necessary for a variety of protein functions including catalysis and regulation. Crystallography and cryoelectron microscopy have provided extensive structural information about local energetic minima in these functional landscapes. Recent experimental advances in techniques such as single molecule Förster resonance energy transfer and nuclear magnetic resonance have shown that proteins and large molecular assemblies are highly dynamic. These motions take place over large-length and long-time scales. While these experimental studies have provided tremendous insights into the functional dynamics of biomolecular systems, computer simulations offer the potential to bridge static structural data with dynamic experiments at atomic resolution.

Consideration of a fundamental dynamic process, folding, has motivated the energy landscape theory of protein folding (1–3). The theory states that evolution has achieved folding robustness by selecting for sequences where the interactions present in the functionally competent states are mutually consistent. The energy landscape for such sequences has an overall funnel shape, which has an enormous influence on folding mechanisms. Computational models that take advantage of the funneled nature of the energy landscape are called 'structure-based models' (SBM) (4,5). The success of SBM and their interplay with experiments has led to a deeper understanding of the underlying physical properties that determine folding dynamics (3).

Since the funneled energy landscape upon which biomolecules fold is the same landscape that governs the functionally important motions, SBM have been used to study long-time and large-length scale molecular dynamics, e.g. (3–10). The simplest varieties of SBM are coarse-grained, where each residue is represented by a single bead and only the interactions present in the native state are attractive (4). All-atom SBM allow a more explicit connection with experimental observables and have been used to understand the interplay between side chain and backbone dynamics during protein and RNA folding (5,6).

The SMOG@ctbp web server is available to facilitate creation and use of SBM to investigate the dynamics of proteins, RNA and DNA. Both C_α (4) and all-atom (5,6) models are available. The SBM represents a baseline model upon which additional complexity can be added by the user. Any PDB structure consisting of any number of chains of standard amino acids, RNA, DNA

*To whom correspondence should be addressed. Tel: +1 858 534 7067; Fax: +1 858 534 7697; Email: jonuchic@ctbp.ucsd.edu

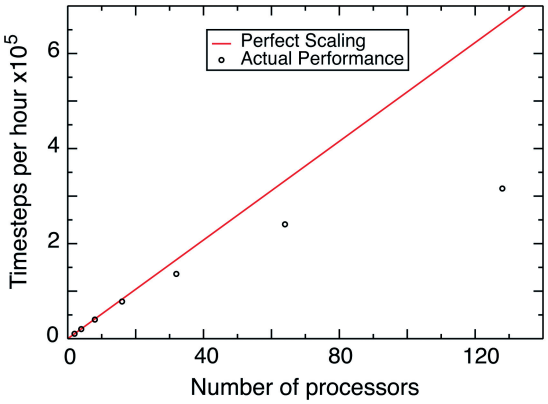

Figure 1. Performance of an all-atom structure-based simulation with GROMACS version 4.0.5 for a ribosome with 142 196 atoms (PDB codes: 2WDG, 2WDH). The system scales up to 32–64 processors before significant performance loss. Due to large amounts of empty space inside the ribosome, this represents a lower bound on potential scalability. See Supplementary Data for more detail.

and a small library of ligands can be directly uploaded to the web server. The necessary files to implement the SBM in GROMACS (11) are provided as output. GROMACS is a state-of-the-art open-source molecular dynamics package that has the flexibility necessary to implement an efficient and highly scalable SBM (Figure 1; [10]). In this article, we describe the structure-based Hamiltonian and web server located at http://smog.ucsd.edu. Further explanation is included in the Supplementary Data and on the web server itself.

METHODS

Formulation of the Hamiltonian

In the creation of GROMACS topology files, the web server uses previously published and validated structure-based Hamiltonians for the C_α (4) and all-atom (5,6) models. The functional form of the C_α Hamiltonian $V_{C\alpha}$ is,

$$V_{C\alpha} = \sum_{\text{bonds}} \varepsilon_r(r-r_o)^2 + \sum_{\text{angles}} \varepsilon_\theta(\theta-\theta_o)^2 + \sum_{\text{backbone}} \varepsilon_D F_D(\phi) + \sum_{\text{contacts}} \varepsilon_C\left[5\left(\frac{\sigma_{ij}}{r}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r}\right)^{10}\right] + \sum_{\text{non-contacts}} \varepsilon_{NC}\left(\frac{\sigma_{NC}}{r}\right)^{12} \quad (1)$$

and the all-atom Hamiltonian V_{AA} is,

$$V_{AA} = \sum_{\text{bonds}} \varepsilon_r(r-r_o)^2 + \sum_{\text{angles}} \varepsilon_\theta(\theta-\theta_o)^2 + \sum_{\text{impropers/planar}} \varepsilon_\chi(\chi-\chi_o)^2 + \sum_{\text{backbone}} \varepsilon_{BB} F_D(\phi) + \sum_{\text{sidechains}} \varepsilon_{SC} F_D(\phi) + \sum_{\text{contacts}} \varepsilon_C\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r}\right)^{6}\right] + \sum_{\text{non-contacts}} \varepsilon_{NC}\left(\frac{\sigma_{NC}}{r}\right)^{12} \quad (2)$$

where the dihedral potential F_D is,

$$F_D(\phi) = [1-\cos(\phi-\phi_o)] + \frac{1}{2}[1-\cos(3(\phi-\phi_o))]. \quad (3)$$

A comprehensive listing and explanation of the parameters is available elsewhere (4–6). Here, we provide an overview of the forcefields for new users of these models. All geometrical parameters are determined from the provided PDB structure, such that the lowest energy value of each term corresponds to the PDB file configuration. Accordingly, by construction, the lowest energy configuration is the PDB structure. If multiple chains are present, any contacts between chains are equal in strength to intra-chain contacts.

The C_α model is defined only for proteins. C_α representations of RNA and DNA are not supported. In the C_α model, each residue is represented by a single bead, located at the position of the C_α atom. Bonds, bond angles and backbone dihedrals are between two, three and four consecutive beads, respectively. Backbone dihedrals and contacts are equally weighted and contacts are defined between native residue contacts.

The all-atom model includes each heavy atom and hydrogens are excluded. Since the atomic geometry is explicitly represented, bonds, bond angles and dihedrals have their traditional meanings. Improper dihedrals are included to preserve chirality, and where necessary, planarity. Contacts are defined between native atom pairs. In contrast to the C_α model, the overall interaction strength between residues is heterogeneous, since residues can have differing numbers native atom–atom pairs. This heterogeneity was shown to have only a weak effect on overall folding mechanisms for small globular proteins (5), though the small differences that arise can increase the agreement between experimental and theoretical ϕ-values (12).

Contact map

The SBM Hamiltonian can be roughly partitioned into two components, local terms to maintain the geometry and local bias, and non-local terms to provide the excluded volume and tertiary bias. The biasing non-local terms are contained within a 'native contact map' and are called 'contacts'. It should be noted that a subset of these contacts which are between atoms close in sequence, in particular 1–5 interactions, contribute to the local bias. Any atoms not interacting through a contact, bond, angle or dihedral, are considered 'non-contacts' and interact only through excluded volume. In the C_α model, the contacts are defined between residue pairs and in the all-atom model between atom pairs. The contacts are determined from the given PDB structure. A pair of residues is defined as being in contact if any shared atom pair is in contact. In this web tool, we allow three possible definitions of native contacts: a 'Cut-off' distance criteria, the 'Shadow' algorithm or 'User Defined'. The cut-off criterion defines two atoms in contact if the atom centers are within 4 Å in the provided PDB structure. The Shadow definition considers all atom pairs within a 6 Å cut-off and then excludes any atom pairs which have an

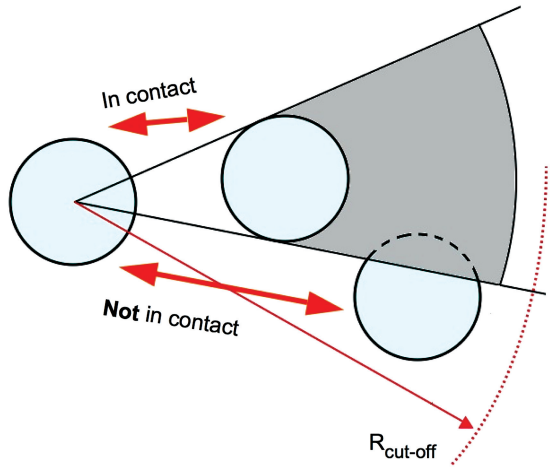

Figure 2. Shadow contact algorithm. To determine the contacts of atom *i*, all atoms within a cut-off radius of atom *i* are considered. The algorithm effectively replaces atom *i* with a light source. Adjacent atoms are represented as opaque spheres with a radius of 1Å. All atoms within the cut-off that have a shadow cast upon them are discarded. The remaining atoms within the cut-off are defined as 'in contact' with atom *i*.

occluding atom (Figure 2). Essentially, Shadow attempts to determine all contacts between interior protein surfaces without allowing atoms to interact through other atoms. The Shadow algorithm is explained in detail on the web server and is the subject of ongoing investigations.

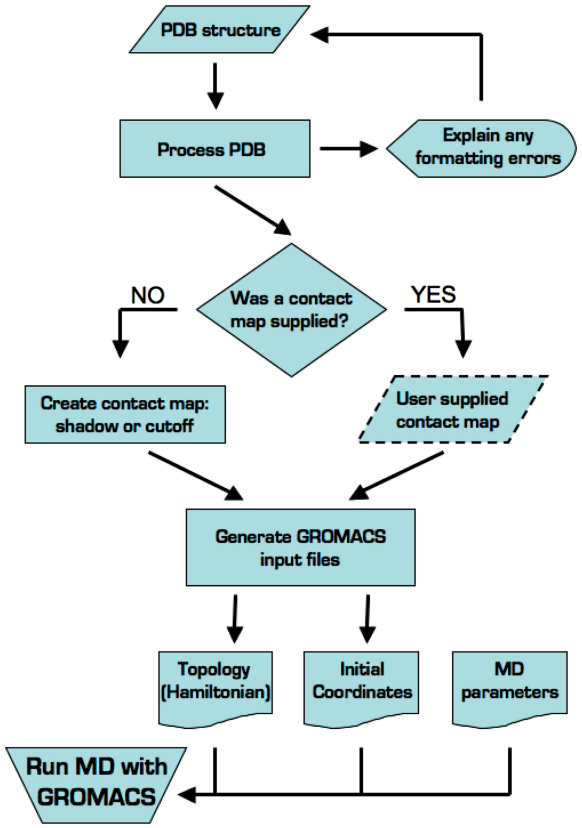

Figure 3. Flowchart explaining the logic of the SMOG@ctbp web server.

IMPLEMENTING A STRUCTURE-BASED MODEL IN GROMACS

Web server interaction

The main purpose of the web server is to create the input files necessary to simulate a biomolecular system with a SBM in GROMACS (Figure 3). A PDB structure that is uploaded from the user's computer is the only required input. While most PDB structures can be directly downloaded from the PDB database and used with the web tool, users should verify that the PDB file conforms to the guidelines described below and in the Supplementary Data. A valid PDB file has a `TER` statement (left justified) in between each chain and an `END` statement (left justified) at the end. The following residues are supported by the web tool:

- Protein residues: all standard 20 amino acids (three letter codes used).
- RNA residues: CYT or C, GUA or G, URA or U and ADE or A.
- DNA residues: DG, DC, DA, DT.
- Ligands: SAM (S-adenosylmethionine), GNP (Gpp (NH)p), ATP, ADP, AMP

Upon request, additional ligands may be supported.

The web page where the PDB file is uploaded is entitled 'Prepare a Simulation' and is where all user input is obtained. Beyond uploading a PDB file, the web server interface allows the user to customize some basic parameters of the SBM Hamiltonian:

(1) The level of graining: it can be varied between all-atom and C_α.
(2) The contact map: the user can upload a native contact map or generate a map by choosing either the cut-off or Shadow algorithm. The contact map algorithms are based on the all-atom geometry, thus PDB files that lack some heavy atoms must be manually inspected to ensure proper performance.
(3) The distribution of stabilizing energy: it can be varied between contacts, backbone dihedrals and side chain dihedrals. This is explored in detail (5).
(4) The size of atoms: this can be controlled through σ_{NC}.
(5) The buffer space: the space between the system and the simulation box is an important parameter. Improved performance and effective parallelization in GROMACS depends on periodic boundary conditions being employed. When using the 'dynamic load balancing' features of GROMACS, excessive volumes of empty space can lead to poor scalability. Though, if the simulation box size is too small, the system can interact with its image. While the default 10Å buffer is sufficient for many simulations, for folding, the box size should be nearly the linear length of the molecule.

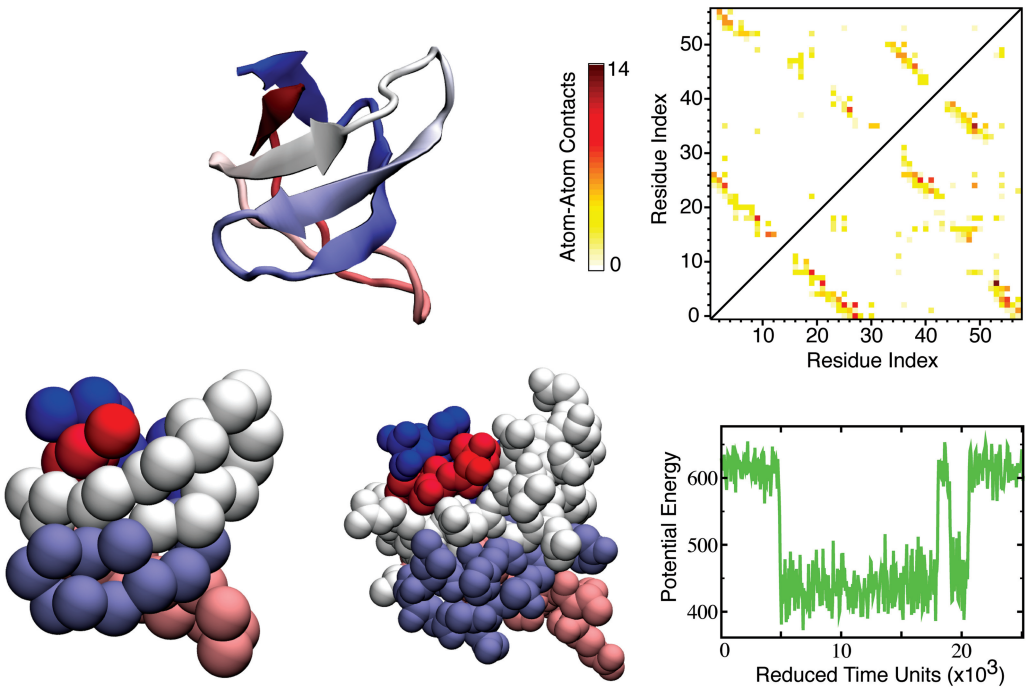

Figure 4. SBM of SH3 domain. PDB code: 1FMK. **Top Left**: Cartoon representation of SH3 domain. **Bottom Left**: C_α model geometry. **Bottom Middle**: All-atom model geometry. **Top Right**: Contact map for SH3. Upper triangle shows 4Å cut-off and lower triangle shows Shadow. Coloring is by number of atom–atom pairs per residue–residue contact. **Bottom Right**: Folding of 57-residue SH3 domain at constant reduced temperature $\tilde{T} = 0.89$ with the all-atom model. Residues 84–140 taken directly from 1FMK.pdb and submitted at SMOG@ctbp with default parameters and Shadow contact map. MD parameters file taken from the web server example.

After uploading a PDB file, inspecting the above parameters, and pressing the 'Submit' button, the web server will either return a link to the completed output or return an error message describing any formatting inconsistencies. The completed output is a tarball containing:

(1) GROMACS coordinate file: the initial structure corresponding to the provided PDB structure; shifted such that the box starts at the origin (`.gro`).
(2) GROMACS topology file: describes all the atomic interactions in the SBM Hamiltonian (`.top`).
(3) GROMACS index file: convenient for manipulating structures with multiple chains (`.ndx`).
(4) Native contact map: if Shadow is selected (`.contact`).
(5) Web server output: contains any non-fatal warnings and messages (`.output`).

Molecular dynamics with GROMACS

In order to run molecular dynamics, the user must have access to a compiled GROMACS 4 distribution. The GROMACS source code can be found at http://www.gromacs.org. The topology file and coordinate file, along with a molecular dynamics parameter settings file (`.mdp`) are sufficient to run the SBM in GROMACS. A suggested `.mdp` is available on the web server. Example output for an SH3 domain is shown in Figure 4. See the web server or the Supplementary Data for a brief tutorial highlighting the relevant GROMACS syntax and things to consider.

CONCLUSION

In this article, we describe SMOG@ctbp, a web server that creates the necessary files to simulate a SBM in GROMACS from a provided PDB structure. The all-atom SBM represents a baseline model that the user is welcome to augment and explore with system-dependent details, e.g. electrostatics or non-native interactions. The possible applications of SBM go beyond equilibrium and kinetic molecular dynamics. A SBM is a starting point for any study where the overall geometry of the biomolecules is maintained, e.g. fitting crystallographic structures into cryoelectron microscopy maps (13) and predicting protein–DNA complexes (14). Hopefully, SMOG@ctbp will enable users to conceive of more new and exciting applications of SBM.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the New Mexico Computing Applications Center for computing time on the Encanto Supercomputer.

FUNDING

Center for Theoretical Biological Physics, National Science Foundation (PHY-0822283, NSF-MCB-0543906); the LANL LDRD program; National Institutes of Health (R01-GM072686); National Institutes of Health Molecular Biophysics Training Program at University of California at San Diego (T32GM08326 to J.K.N.). P.C.W. is currently funded by a LANL Director's Fellowship. Funding for open access charge: Center for Theoretical Biological Physics at UCSD.

Conflict of interest statement. None declared.

REFERENCES

1. Bryngelson,J.D. and Wolynes,P.G. (1987) Spin-glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.
2. Leopold,P.E., Montal,M. and Onuchic,J.N. (1992) Protein folding funnels - a kinetic approach to the sequence structure relationship. *Proc. Natl Acad. Sci. USA*, **18**, 8721–8725.
3. Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.
4. Clementi,C., Nymeyer,H. and Onuchic,J.N. (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.*, **298**, 937–953.
5. Whitford,P.C., Noel,J.K., Gosavi,S., Schug,A., Sanbonmatsu,K.Y. and Onuchic,J.N. (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins: Struct. Funct. Bioinf.*, **75**, 430–441.
6. Whitford,P.C., Schug,A., Saunders,J., Hennelly,S.P., Onuchic,J.N. and Sanbonmatsu,K.Y. (2009) Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function. *Biophys. J.*, **96**, L7–L9.
7. Whitford,P.C., Miyashita,O., Levy,Y. and Onuchic,J.N. (2007) Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.*, **366**, 1661–1671.
8. Hyeon,C. and Onuchic,J.N. (2007) Mechanical control of the directional stepping dynamics of the kinesin motor. *Proc. Natl Acad. Sci. USA*, **104**, 17382–17387.
9. Pincus,D.L., Cho,S.S., Hyeon,C.B. and Thirumalai,D. (2008) Minimal models for proteins and RNA: from folding to function. *Prog. Mol. Biol. Transl. Sci.*, **84**, 203–250.
10. Whitford,P.C., Geggier,P., Altman,R.B., Blanchard,S.C., Onuchic,J.N. and Sanbonmatsu,K.Y. (2010) Accommodation of aminoacyl-tRNA into the ribosome involves reversible excursions along multiple pathways. *(RNA)*, **16**, 1196–1204.
11. Hess,B., Kutzner,C., van der Spoel,D. and Lindahl,E. (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.
12. Clementi,C., Garca,A.E. and Onuchic,J.N. (2003) Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.*, **326**, 933–954.
13. Orzechowski,M. and Tama,F. (2008) Flexible fitting of high-resolution X-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.*, **95**, 5692–5705.
14. Schug,A., Weigt,M., Onuchic,J.N., Hwa,T. and Szurmant,H. (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl Acad. Sci. USA*, **106**, 22124–22129.