BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 21 2013, pages 2795–2796
doi:10.1093/bioinformatics/btt478

*Structural bioinformatics*

Advance Access publication September 9, 2013

# eSBMTools 1.0: enhanced native structure-based modeling tools

Benjamin Lutz[1,2], Claude Sinner[1,2], Geertje Heuermann[1,2], Abhinav Verma[2] and Alexander Schug[2,*]

[1]Department of Physics and [2]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany

Associate Editor: Anna Tramontano

**ABSTRACT**

**Motivation**: Molecular dynamics simulations provide detailed insights into the structure and function of biomolecular systems. Thus, they complement experimental measurements by giving access to experimentally inaccessible regimes. Among the different molecular dynamics techniques, native structure-based models (SBMs) are based on energy landscape theory and the principle of minimal frustration. Typically used in protein and RNA folding simulations, they coarse-grain the biomolecular system and/or simplify the Hamiltonian resulting in modest computational requirements while achieving high agreement with experimental data. eSBMTools streamlines running and evaluating SBM in a comprehensive package and offers high flexibility in adding experimental- or bioinformatics-derived restraints.

**Results:** We present a software package that allows setting up, modifying and evaluating SBM for both RNA and proteins. The implemented workflows include predicting protein complexes based on bioinformatics-derived inter-protein contact information, a standardized setup of protein folding simulations based on the common PDB format, calculating reaction coordinates and evaluating the simulation by free-energy calculations with weighted histogram analysis method or by phi-values. The modules interface with the molecular dynamics simulation program GROMACS. The package is open source and written in architecture-independent Python2.

**Availability:** http://sourceforge.net/projects/esbmtools/

**Contact:** alexander.schug@kit.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on May 23, 2013; revised on July 22, 2013; accepted on August 9, 2013

## 1 INTRODUCTION

Energy landscape theory and the principle of minimal frustration stipulate that biomolecular evolution results in an effective overall bias of a macromolecule's energy landscape being funnel-shaped toward its global minimum at the native folded state (Onuchic and Wolynes, 2004; Schug and Onuchic, 2010). Native structure-based models (SBMs, also often referred to as Gō-models) realize an idealized minimally frustrated folding funnel devoid of any energetic traps by a simple Hamiltonian (Clementi *et al*., 2000; Whitford *et al*., 2009) that allows molecular dynamics simulations to reach biologically relevant timescale.

*To whom correspondence should be addressed.

The SBM potential is based on the native folded structure and native contact information. The number of formed native contacts in a structure is the dominant reaction coordinate typically called $Q$. The $Q$ value can be used as a reaction coordinate to investigate folding paths or as an order parameter for the weighted histogram analysis method (Kumar *et al*., 1992).

The SBM can be combined with coarse-graining approaches. Proteins can be represented by single beads at the positions of their $C_\alpha$ atoms. Accordingly, there exist formulations of the SBM at an all-atom (Whitford et al., 2009) and $C_\alpha$ (Clementi *et al*., 2000) graining level (see Supplementary Information). Bioinformatics-derived contact information (Schug *et al*., 2009) or experimental measurements (Whitford *et al*., 2011) can be added as restraints to SBM.

Eventually, the SBM defines a force field that results in computationally tractable simulations that facilitate simulations on the effective timescale of seconds (for small proteins even on single CPUs) in combination with extensive sampling. Existing web-based solutions like the SMOG-server (Noel *et al*., 2010) allow a straightforward setup and evaluation of standard SBM. Our open source eSBMTools extends the customization options enabling the setup of automatized workflows and strong modifications of SBM by, e.g. adding arbitrary additional contacts, protein/RNA structures, novel ligand topologies or manipulating force field parameters.

## 2 FUNCTIONALITY

eSBMTools is organized in modules that can be loaded into Python projects. The modules and their input and output files interface with tools that are part of the standard GROMACS software package (Hess *et al*., 2008). It can be used at all stages in the context of SBM simulations: from generating the SBM itself, over manipulations of the model and configuration file generation, to extensive post-processing of simulation data. Several examples can be found in the Supplementary Information.

The SBM is generated starting from a PDB structure with the help of an XML-based topology definition that introduced the bonded interactions via bonds, angles and dihedral angles. eSBMTools provides topology definitions for DNA/RNA or protein systems in an all-atom or a $C_\alpha$ formulation. The XML format allows the user to easily extend the existing topology definitions by custom additions of arbitrary ligands or experimental markers (for example FRET-fluorosphores). The non-bonded interactions via inter-atomic contacts are implemented by a simple cut-off formulation. Two atoms are considered to form a contact when the distance between the atoms is below

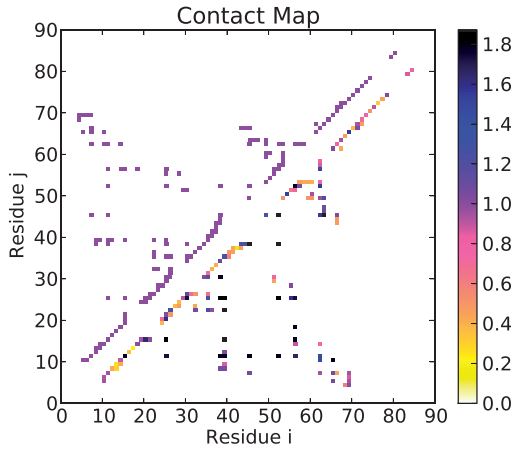


**Fig. 1.** Contact maps of protein *HigA* in the standard homogeneous formulation and the Miyazawa–Jernigan formulation. Each square stands for a contact between residue $i$ and $j$ in the native structure of *HigA*. The upper left contact map represents the homogeneous energetics of the standard SBM formulation. However, the lower right contact map illustrates the possibility of weighting each native contact by Miyazawa–Jernigan factors (Miyazawa and Jernigan, 1996)

a certain threshold. A contact map represents every contact between two atoms by an entry, as in Figure 1.

The generated SBM can be modified according to the requirements of the user's system of interest. The contact information can be manipulated to include, e.g. predicted contacts (Dago *et al.*, 2012; Morcos *et al.*, 2011) or the energetics of all contacts can be re-weighted by, e.g. by amino acid interaction matrices (Miyazawa and Jernigan, 1996), shown in Figure 1.

Two SBMs can be merged to set up combined systems for, e.g. complex formation simulations (Schug *et al.*, 2009). Once all the files that describe the system of interest are prepared, eSBMTools generates the required configuration files for GROMACS simulations. This generation can be customized within Python to adjust simulation parameters, e.g. temperature, duration or resolution to steer simulations within a workflow.

The simulation output by GROMACS can be processed in various ways depending on the desired evaluation protocol. The root-mean-square deviations along the trajectory of the simulation can be generated and included in further post-processing. Similarly, the $Q$ values (the number of formed native contacts) can be generated from a trajectory. The $Q$ values can be used to generate time-depending contact maps or filtered by certain ranges to follow the folding process of sub-structures within the system of interest. Based on histograms of $Q$ values, it is possible to calculate $\phi$ values (Fersht and Sato, 2004) that give a measure for a residue's stability in the transition state. eSBMTools can also be used to perform a more demanding analysis of $Q$ value and energy histograms: the weighted histogram analysis method (Kumar *et al.*, 1992). This method yields the free-energy landscape $\Delta G(Q, T)$ over the sampled $Q$ values and a chosen temperature range, based on a finite number of simulations at distinct temperatures, preferably around the folding temperature. In addition, it is possible to calculate the heat capacity over the temperature range, which allows the user to determine the folding temperature.

## 3 CONCLUSIONS

We present the full release version of our Python2-based software package for SBM simulation pre- and post-processing. The package provides a comprehensive tool box to perform and evaluate simulations in the field of SBMs. The open source package interfaces with a standard build of the GROMACS software suite, which is a state-of-the-art software for molecular dynamics simulations. The combination of a high-performance simulation software with a user-defined Python2 framework makes eSBMTools versatile yet compact. Computationally tractable SBM can be set up in a platform-independent manner as workflows for extensive sampling on both HPC systems and local computers.

*Funding*: AS recognizes support from the Helmholtz Association within the 'Impuls- und Vernetzungsfond'.

*Conflict of interest*: none declared.

## REFERENCES

Clementi,C. *et al.* (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.

Dago,A.E. *et al.* (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc. Natl Acad. Sci. USA*, **109**, E1733–E1742.

Fersht,A.R. and Sato,S. (2004) Phi-value analysis and the nature of protein-folding transition states. *Proc. Natl Acad. Sci. USA*, **101**, 7976–7981.

Hess,B. *et al.* (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.

Kumar,S. *et al.* (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, **13**, 1011–1021.

Miyazawa,S. and Jernigan,R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA*, **108**, E1293–E1301.

Noel,J.K. *et al.* (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.*, **38**, W657–W661.

Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.

Schug,A. and Onuchic,J.N. (2010) From protein folding to protein function and biomolecular binding by energy landscape theory. *Curr. Opin. Pharmacol.*, **10**, 709–714.

Schug,A. *et al.* (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl Acad. Sci. USA*, **106**, 22124–22129.

Whitford,P.C. *et al.* (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins*, **75**, 430–441.

Whitford,P.C. *et al.* (2011) Excited states of ribosome translocation revealed through integrative molecular modeling. *Proc. Natl Acad. Sci. USA*, **108**, 18943–18948.