Ensembler: Enabling high-throughput molecular simulations at the superfamily scale

Daniel L. Parton,[1] Patrick B. Grinaway,[1] and John D. Chodera[1,*]

[1]*Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065*

(Dated: January 23, 2015)

The rapidly expanding body of available genomic and protein structural data provides a rich resource for the field of biomolecular simulation. However, simulations on an omics scale are not yet widely performed, partly because software has had trouble keeping pace. For example, it should be possible to study proteins across entire (super)families, and to do so in a way which exploits the entire variety of available structural biology data. Here, we present a new tool for enabling high-throughput simulation in the genomics era. Ensembler takes any set of sequences - from a single sequence to an entire superfamily of interest - and shepherds them through various stages of: comparative modeling to all relevant PDB structures, reconstruction of missing loops, addition of missing atoms, culling by close structural similarity, assignment of protonation states, solvation, and refinement with molecular simulation. The output is an ensemble of structures ready for subsequent parallel or distributed molecular simulations. This automates much of the time-consuming process of preparing protein models suitable for simulation, while also allowing this process to be scaled to the superfamily scale. A particular advantage of this approach can be found in the construction of kinetic models of conformational dynamics - such as Markov state models - for which a diverse array of starting configurations is expected to aid sampling. We demonstrate the power of this approach by constructing initial models for all catalytic domains in the human tyrosine kinase family, using all kinase catalytic domain structures from any organism as structural templates. Ensembler should run on all major operating systems, and has been tested on Linux and OS X. The program is free of charge, and is made available under the terms of the GNU General Public License (GPL) v2. The latest release can be installed via the conda package manager, and the latest source can be downloaded from https://github.com/choderalab/ensembler.

I. INTRODUCTION

II. DESIGN AND IMPLEMENTATION

III. RESULTS

IV. AVAILABILITY AND FUTURE DIRECTIONS

V. ACKNOWLEDGMENTS

* Corresponding author; john.chodera@choderalab.org