Memorial Sloan-Kettering
Cancer Center

Dr. John Chodera
Assistant Member
Computational Biology Center
Sloan-Kettering Institute

23 Jun 2015

To:

Editors,
PLoS Computational Biology

Re: PRE-SUBMISSION INQUIRY FOR SOFTWARE PAPER SUBMISSION

Dear Editors:

We wish to submit for your consideration a software paper entitled:

Ensembler: Enabling high-throughput molecular simulations at the superfamily scale.
by Daniel L Parton, Patrick B Grinaway, Sonya M Hanson, Kyle A Beauchamp, John D Chodera.

Ensembler is a software pipeline for automated, high-throughput protein modeling and biomolecular simulation preparation intended to enable the field of biomolecular simulation to move into simulating proteins at the family or superfamily scale. **Ensembler** addresses two major limitations in the biomolecular simulation field, by a) automating the time-consuming procedures typically required for simulation setup, which otherwise restrict most simulation studies to one or a few proteins, and b) by allowing a user to exploit the full variety of available genomic and structural data to generate configurationally diverse arrays of protein conformations for one target protein, a few target proteins, or entire (super)families of proteins. These models can be subjected to immediate structural analysis or used as starting conditions for highly parallel molecular dynamics simulations, which are increasingly becoming commonplace thanks to the availability of inexpensive graphics processing units (GPUs). In the latter case, the diversity of structures generated by **Ensembler** – which may include a variety of structures of potential functional relevance – is expected to greatly aid configurational sampling, especially in the use of state-of-the-art methods for constructing kinetic models of protein dynamics, such as Markov state models, that aggregate data from multiple independent simulation trajectories.

To illustrate the utility of **Ensembler**, we have applied it to the set of known human tyrosine kinase domains. By using all structures of protein kinase domains (of any species) as templates, over 4,000 models are generated for each tyrosine kinase domain

1275 York Avenie New York, NY 10065
Phone: 646-888-3400 Fax: 646-888-3105
choderaj@mskcc.org Choderalab.org

Memorial Sloan-Kettering
Cancer Center

Dr. John Chodera
Assistant Member
Computational Biology Center
Sloan-Kettering Institute

target sequence. We show that these models cover a broad range of configurations, and, in the specific cases of Src and Abl1, include states known to be involved in their activation mechanisms, with many configurations in-between. The models are thus expected to greatly aid sampling in highly parallel molecular dynamics simulations, which we plan to use to generate Markov state models to understand and describe the conformational landscape of the kinase family.

We expect this software to be useful to a broad community of users, including groups who want to use **Ensembler** models to seed highly parallel molecular dynamics simulations of single proteins or large protein families, as well as those who want to use them to explore the configurational diversity of a given protein family. The source code is written Python and is freely available at GitHub. Pre-packaged Python binaries can also easily be automatically installed using the conda package manager for Python. The code is licensed under the open-source GNU General Public License, version 2 (GPLv2). Full documentation is provided at ensembler.readthedocs.org.

There appears to be significant interest in this work already. Drafts of the manuscript describing **Ensembler** that we hope to submit to PLoS Computational Biology have been available on bioRxiv ahead of submission, at

http://biorxiv.org/content/early/2015/06/29/018036

and the manuscript PDF has been downloaded over 339 times since it was posted. Furthermore, the source code has been downloaded from GitHub 76 times, by 27 unique users.

Sincerely,

John D. Chodera

Assistant Member, Computational Biology Program
Sloan-Kettering Institute, Memorial Sloan-Kettering Cancer Center

1275 York Avenie New York, NY 10065
Phone: 646-888-3400 Fax: 646-888-3105
choderaj@mskcc.org Choderalab.org

Memorial Sloan-Kettering
Cancer Center

Dr. John Chodera
Assistant Member
Computational Biology Center
Sloan-Kettering Institute

Assistant Professor of Physiology, Biophysics and Systems Biology Program
Weill Cornell Graduate School of Medical Sciences

Phone: 646.888.3400
Email: john.chodera@choderalab.org

1275 York Avenie New York, NY 10065
Phone: 646-888-3400 Fax: 646-888-3105
choderaj@mskcc.org Choderalab.org