Review Criteria:

- *Originality*: This is a highly original work to address an important unsolved problem in computational molecular biophysics, namely how to leverage the wealth of structural data in the Protein Data Bank to seed simulations used to create Markov State Models (MSMs).
- *High importance to researchers in computational biology*: With increasing computational power of multicore CPUs, coprocessors and distributed computing, it is possible to construct MSMs, which are kinetic models of conformational dynamics. Creation of MSMs is highly accelerated by seeding with a diverse population of conformations, which is elegantly addressed by the Ensembler program.
- *Significant biological insight and general interest to life scientists*: Simulation plays an important role in complementing traditional experimental methods such as X-ray crystallography, whose structures are typically limited to fluctuations about a single conformational state. Ensembler enables conformational states discovered for homologous sequences using traditional experimental methods to be leveraged as seeds for MSM construction for the sequence of interest. This approach opens the door to biological insights that previously would necessitate significant computational expense due to the availability of fewer conformational seeds. The biological insights presented for SRC and ABL are limited, but that's not the point of this work; the point is to generate starting conformations for building MSMs that *will* provide exciting SRC/ABL insights difficult or impossible to capture experimentally.
- *Rigorous methodology:* In Ensembler, the authors put forward a reasonably flexible implementation that allows the user some control in choosing the level of rigor (i.e. the force field and range of templates). One concern is that its difficult to ascertain how many of the generated models (i.e. 4003 for SRC and 4280 for ABL) are truly viable starting conformations for production simulations, or actually represent conformations so high in energy they should be pruned. We recommend adding 1 layer to the protocol based on evaluating the MolProbity score of each generated model. For example, the Python implementation of MolProbity in Phenix could be used.
- *Substantial evidence for its conclusions*: The major conclusion of this work is that the Ensembler program streamlines creation of production ready simulation systems for a target sequence(s), which span all available (or a selected subset) of experimental models for a chosen protein family. This is an exciting advance, which has the potential to greatly accelerate generation of MSMs. However, sometimes less is more. The evidence presented in the paper is not at all convincing that the ~4000 models generated for SRC or ABL are actually worthy of further computational effort. This may cause significant problems for users of Ensembler; for example, how many of the SRC/ABL confirmations are so strained that they will unfold upon simulations longer than a few nanoseconds. We suggest the authors evaluate all models using the MolProbity tool to uncover the number of vdW clashes, poor side-chains rotamers, poor backbone conformations, etc., as a way to allow users to filter out unproductive conformations.

Major Issues:

As discussed above, the only major issue is that it is unclear how many of the generated simulation systems are too strained to be worthy of further computational resources. This does not seem to be consistent with “best practices”. We suggest a cheap MolProbity evaluation be run on each structure as a way to truly establish quality criteria users can rely on. This may save users of Ensembler the pain and frustration of simulating highly strained structures that aren’t actually generating relevant conformational statistics.

Minor issues:

Line 201: “guaranteeing some degree of homology between targets and templates.”
Sequences are either homologous (i.e. there is an evolutionary relationship), or they are not homologous. The phrase “degree of homology” is not meaningful.
A suggested rephrasing is “guaranteeing homology and some degree of sequence identity.”

Line 208: “e.g. for crystal unit cells with multiple asymmetric units.”
All unit cells have a single asymmetric unit. The phase multiple asymmetric units is not meaningful.
A suggested rephrasing is “for crystals with non-crystallographic symmetry giving rise to independent conformations of the sequence within the asymmetric unit”.

Line 288: “with the PAM 250 scoring matrix of Gonnet et al.”
Why PAM250 rather than a BLOSUM scoring matrix?