idpflex: Analysis of Intrinsically Disordered Proteins by Comparing Simulations to Small Angle Scattering Experiments

Jose M. Borreguero[1], **Fahima Islam**[1], **Utsab R. Shrestha**[2], **and Loukas Petridis**[2]

1 Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge TN, USA **2** Biosciences Division, Oak Ridge National Laboratory, Oak Ridge TN, USA.

DOI: 10.21105/joss.01007

Software

- Review
- Repository
- Archive

Submitted: 15 August 2018
Published: 10 December 2018

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

It is estimated that about 30% of the eucariotic proteome consists of intrinsically disordered proteins (IDP's), yet their presence in public structural databases is severely underrepresented. IDP's adopt heterogeneous inter-converting conformations with similar probabilities, preventing resolution of structures with X-Ray diffraction techniques. Small angle scattering (SAS) probes the average features of the conformational ensemble, which can prove unsatisfactory when very different ensembles share nearly identical average features. To address this shortcoming, atomistic molecular dynamics (MD) simulations combined with enhanced sampling methods such as the Hamiltonian replica exchange method are specially suitable (al, 2006), since they probe extensive regions of the IDP's free-energy phase space. Hence, this type of simulations can produce physically meaningful conformations and offer a full-featured description of the conformational landscape when properly validated against available experimental SAS data. The python package idpflex clusters the 3D conformations resulting from an MD simulation into a hierarchical tree, with protein substates taking the role of tree nodes. Alternatively, the conformational ensemble of structures can be generated by other means than MD, such as torsional sampling of the protein backbone (J. E. C. et al., 2012). This flexibility is possible because idpflex is agnostic to the ensemble generation method. In contrast to other methods (B. R. et al., 2011), idpflex does not initially discard any conformation by labelling it as incompatible with the available experimental data. The data represent an average over all conformations, and using an average as the criterion by which to discard any specific conformation can lead to erroneous discarding decisions. Clustering is performed according to structural similarity between conformations, defined by the root mean square deviation algorithm (Kabsch, 1976). Alternatively, idpflex can cluster conformations according to an Euclidean distance in the space spanned by a set of structural properties such as radius of gyration and end-to-end distance. Calculation of SAS intensities (D. S. et al., 1995) for each substate allows quantitative comparison to SAS data, yielding the probability of the IDP to adopt one of the conformations of any specific substate. Arranging tens of thousands of conformations into (typically) less than ten substates provides the researcher with a manageable number of macro-conformations from which to derive meaningful conclusions regarding the conformational freedom of the IDP. In addition to clustering, idpflex can compute structural features for each substate such as radius of gyration, end-to-end distance, asphericity, solvent exposed surface area, contact maps, and secondary structure content. All these properties require atomistic detail, thus idpflex is more apt for the study of IDP's than for the study of quaternary protein

arrangements, where clustering of coarse-grain simulations becomes a better option (B. R. et al., 2011). In summary, idpflex integrates MD simulations with SAS experiments to obtain the conformational ensemble of IDP's, a pertinent problem in structural biology.

It is estimated that about 30% of the eucariotic proteome consists of intrinsically disordered proteins (IDP's), yet their presence in public structural databases is severely underrepresented. IDP's adopt heterogeneous inter-converting conformations with similar probabilities, preventing resolution of structures with X-Ray diffraction techniques. An alternative technique with wide application on IDP systems is small angle scattering (SAS). SAS can measure average structural features of IDP's when in vitro solution, or even at conditions mimicking protein concentrations found in the cell's cytoplasm.

Despite these advantages, the averaging nature of SAS measurements will prove unsatisfactory if one aims to differentiate among the different conformations that a particular IDP can adopt. Different distributions of conformations can yield the same average therefore it is not possible to retrace the true distribution if all that SAS provides is the average conformation.

To address this shortcoming, atomistic molecular dynamics (MD) simulations of IDP systems combined with enhanced sampling methods such as the Hamiltonian replica exchange method are specially suitable (al, 2006). These simulations can probe extensive regions of the IDP's conformational space and have the potential to offer a full-featured description of the conformational landscape of IDP's. The results of these simulations should not be taken at faith value, however. First, a proper comparison against available experimental SAS data is a must. This validation step is the requirement that prompted the development of `idpflex`.

The python package `idpflex` clusters the 3D conformations resulting from an MD simulation into a hierarchical tree by means of structural similarity among pairs of conformations. The conformations produced by the simulation take the role of Leafs in the hierarchichal tree. Nodes in the tree take the rol of IDP substates, with conformations under a particular Node making up one substate. Strictly speaking, `idfplex` does not require the IDP conformations to be produced by an MD simulation. Alternative conformation generators can be used, such as torsional sampling of the protein backbone (J. E. C. et al., 2012). In contrast to other methods (B. R. et al., 2011), `idpflex` does not initially discard any conformation by labelling it as incompatible with the experimental data. This data is an average over all conformations, and using this average as the criterion by which to discard any specific conformation can lead to erroneous discarding decisions by the reasons stated above.

Default clustering is performed according to structural similarity between pairs of conformations, defined by the root mean square deviation algorithm (Kabsch, 1976). Alternatively, `idpflex` can cluster conformations according to an Euclidean distance in an abstract space spanned by a set of structural properties, such as radius of gyration and end-to-end distance. Comparison to experimental SAS data is carried out first by calculating the SAS intensities (D. S. et al., 1995) for each conformation produced by the MD simulation. This result in SAS intensities for each Leaf in the hierarchical tree. Intensities are then propagated up the hierarchical tree, yielding a SAS intensity for each Node. Because each Node takes the role of a conformational substate, we obtain SAS intensities for each substate. `idpflex` can compare the SAS intensity of each substate against the experimental SAS data. Also, it can average intensities from different substates and compare against experimental SAS data. The fitting functionality included in `idpflex` allows for selection of the set of substates that will yield maximal similarity between computed and experimental SAS intensities. Thus, arranging tens of thousands of conformations into (typically) less than ten substates provides the researcher with a manageable number of conformations from which to derive meaningful conclusions regarding the conformational variability of IDP's.

`idpflex` also provides a set of convenience functions to compute structural features of IDP's for each of the conformations produced by the MD simulation. These properties can then be propagated up the hierarchical tree much in the same way as SAS intensities are propagated. Thus, one can compute for each substate properties such as radius of gyration, end-to-end distance, asphericity, solvent exposed surface area, contact maps, and secondary structure content. All these structural properties require atomistic detail, thus `idpflex` is more apt for the study of IDP's than for the study of quaternary protein arrangements, where clustering of coarse-grain simulations becomes a better option (B. R. et al., 2011). `idpflex` wraps other python packages (MDAnalysis (N. M.-A. et al., {2011}), (R. J. G. et al., {2016}), mdtraj (R. T. M. et al., {2015})) and third party applications (CRYSOL (D. S. et al., 1995), DSSP (Wolfgang Kabsch & Sander, {1983})) that actually carry out the calculation of said properties. Additional properties can be incorporated by inheriting from the base Property classes.

To summarize, `idpflex` integrates MD simulations with SAS experiments in order to obtain a manageable representation of the rich conformational diversity of IDP's, a pertinent problem in structural biology.

The "notebooks" directory within the source contains two Jupyter notebooks that illustrate the use of idpflex when clustering an example MD trajectory.

Notice of Copyright

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Acknowledgements

This work is sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle LLC, for DOE. Part of this research is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, User Facilities under contract number DE-AC05-00OR22725.

References

al, R. A. et. (2006). A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *Journal of Chemical Theory and Computation*, *2*(2), 217–228. doi:10.1021/ct050250b

al., B. R. et. (2011). SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure*, *19*(1), 109–116. doi:10.1016/j.str.2010.10.006

al., D. S. et. (1995). CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography*, *28*(6), 768–773. doi:10.1107/S0021889895007047

al., J. E. C. et. (2012). SASSIE: A program to study intrinsically disordered biological molecules and macromolecular ensembles using experimental scattering restraints. *Computer Physics Communications*, *183*(2), 382–389. doi:10.1016/j.cpc.2011.09.010

al., N. M.-A. et. ({2011}). MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of Computational Chemistry*, *32*, 2319–2327. doi:{10.1002/jcc.21787}

al., R. J. G. et. ({2016}). MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. *Proceedings of the 15th Python in Science Conference*, 98–105.

al., R. T. M. et. ({2015}). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *BIOPHYSICAL JOURNAL*, *109*(8), 1528–1532. doi:{10.1016/j.bpj.2015.08.015}

Kabsch, W. (1976). Solution for best rotatation to relate 2 sets of vectors. *Acta Crystallograpica Section A*, *32*(SEP1), 922–923. doi:10.1107/S0567739476001873

Wolfgang Kabsch, & Sander, C. ({1983}). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. doi:{10.1002/bip.360221211}