Bioinformatics, 32(16), 2016, 2545–2547
doi: 10.1093/bioinformatics/btw200
Advance Access Publication Date: 19 April 2016
Applications Note

OXFORD

Structural bioinformatics

Mollack: a web server for the automated creation of conformational ensembles for intrinsically disordered proteins

Zachary Ziegler[1,2], Molly Schmidt[2,3], Thomas Gurry[2,4], Virginia Burger[2] and Collin M. Stultz[2,3,4,5,*]

[1]Cornell University, Ithaca, NY 14850, USA, [2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA, [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA, [4]Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA and [5]The Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on October 20, 2015; revised on March 19, 2016; accepted on April 9, 2016

Abstract

Summary: Intrinsically disordered proteins (IDPs) play central roles in many biological processes. Consequently, an accurate description of the disordered state is an important step towards a comprehensive understanding of a number of important biological functions. In this work we describe a new web server, Mollack, for the automated construction of unfolded ensembles that uses both experimental and molecular simulation data to construct models for the unfolded state. An important aspect of the method is that it calculates a quantitative estimate of the uncertainty in the constructed ensemble, thereby providing an objective measure of the quality of the final model. Overall, Mollack facilitates structure-function studies of disordered proteins.
Availability and Implementation: http://cmstultz-mollack.mit.edu
Contact: cmstultz@mit.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Intrinsically disordered proteins (IDPs) constitute a class of biopolymers that sample a diverse set of conformations during their biological lifetime. Constructing a structural model for an IDP requires that one specify a representative set of structures that capture, at a low resolution, the dominant thermally accessible states of the protein. IDP Ensemble construction, albeit challenging, is an essential step towards a comprehensive understanding of the function of these biopolymers (Tompa and Varadi, 2014).

When experimental data on the IDP of interest are available, the ensemble construction problem can be posed as a straightforward optimization problem; i.e. building an ensemble is equivalent to generating a set of structures that yield calculated ensemble averages that agree with experiment. Indeed, a number of ensemble construction algorithms have been developed that use this guiding principle to optimize the choice of structures (Daughdrill *et al.*, 2012; Fu *et al.*, 2014; Jensen *et al.*, 2010; Marsh and Forman-Kay, 2009), and an online database exists for cataloging and storing these data (Varadi *et al.*, 2014). Despite these advances it is important to recognize that the ensemble construction problem itself is inherently degenerate because the number of degrees of freedom in the protein is generally much larger than the number of experimental constraints (Fisher *et al.*, 2010). A consequence of this is that there may be several different ensembles that have calculated ensemble averages that agree with experiment. Therefore agreement with experiment is insufficient by itself to ensure that the resulting ensemble is correct.

In a previous work, we described an ensemble construction algorithm, based on a Bayesian formalism, that generates an ensemble

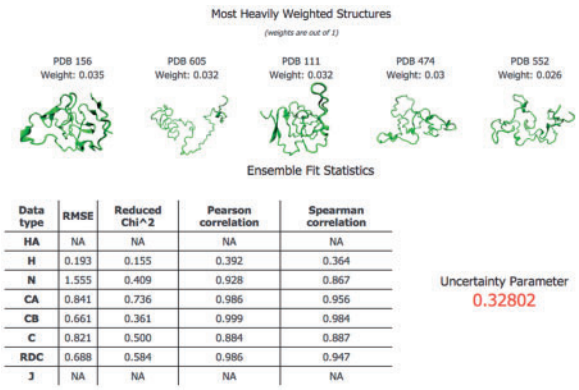
that agrees with a pre-specified set of experimental data (Fisher *et al.*, 2010, 2012). A key aspect of the method is that it also calculates an 'uncertainty parameter', $0 \leq \sigma \leq 1$, which quantifies the uncertainty in the underlying ensemble. When $\sigma = 0$, it is likely that the model is correct, and when $\sigma = 1$ one has little certainty that the constructed ensemble is correct. Nevertheless, when $\sigma \neq 0$ one can generate confidence intervals for ensemble average quantities that provide a rigorous framework for hypothesis testing. Mollack is an online implementation of this Bayesian formalism for constructing ensembles for IDPs.

2 The Mollack Server

The ensemble construction process begins with a structural library consisting of a set of energetically favorable conformations that capture, albeit at low resolution, dominant thermally accessible states of the protein. Since a number of methods exist for generating models of the unfolded state (Jha *et al.*, 2005; Marsh and Forman-Kay, 2009; Ozenne *et al.*, 2012), users are encouraged to upload a pre-generated structural library. However, the user must ensure that the associated structure files are in pdb format and that the structures do not contain bad contacts. Alternatively, users can ask Mollack to generate a library of low energy conformations using only the amino acid sequence of the protein using either GROMACS (Berendsen *et al.*, 1995) or CHARMM (Brooks *et al.*, 2009)—although an academic CHARMM license is needed to use the latter algorithm.

If the user chooses to have Mollack generate their structural library, the algorithm will run replica-exchange simulations, using implicit solvent, to generate a diverse set of low energy structures. Default settings utilize a Generalized Born implicit solvent model (Bashford and Case, 2000) with GROMACS and EEF1 (Lazaridis and Karplus, 1999) with CHARMM. In addition, CHARMM simulations can also employ a biasing potential to ensure that the protein samples states that have different amounts of β-sheet and α-helix content (Gurry and Stultz, 2014). A user can generate different structural libraries by changing the simulation protocol using the advanced options tab. Generating a structural library is the most labor-intensive part of a Mollack run. GROMACS generation takes 20–80 h, depending on the size of the protein (~20 to 150 residues) and CHARMM generation, using the standard parameters, takes 5–20 h. While both simulation methods generate qualitatively similar results, the main difference between the two approaches is that CHARMM runs require less time. This is largely due to the fact that runs with the EEF1 implicit solvent model are faster than runs with the Generalized Born model. Currently, automated structure generation is limited to proteins that are at most 150 residues and runs that last at most 5-days. If more computational resources are required, the user should pre-generate a structural library using another online tool (mentioned above) and upload those data to Mollack.

Once the structural library is specified, Mollack uses a user-generated experimental data, to assign weights to the resulting structures. The experimental data file contains all the experimental measurements, organized by residue, that Mollack will use to build the ensemble. Currently the method can utilize NMR chemical shifts, J-couplings, and residual dipolar couplings (RDCs) for ensemble construction. If information is available on the average radius of gyration, this can be utilized as well. A description of the correct format to use is found on the Data page of the website, as well as an option to either build a blank template from the user's primary sequence or build a template that incorporates NMR data from the



Data type	RMSE	Reduced Chi^2	Pearson correlation	Spearman correlation
HA	NA	NA	NA	NA
H	0.193	0.155	0.392	0.364
N	1.555	0.409	0.928	0.867
CA	0.841	0.736	0.986	0.956
CB	0.661	0.361	0.999	0.984
C	0.821	0.500	0.884	0.887
RDC	0.688	0.584	0.986	0.947
J	NA	NA	NA	NA

Fig. 1. Sample Mollack output

Biological magnetic Resonance Data Bank (Ulrich *et al.*, 2008). Example experimental data files are also provided in the Supplementary data. Future updates to the website will expand the set of available data that the method can exploit.

Mollack clusters structures from the generated structural library to reduce the overall ensemble size down to a manageable number (e.g. typically between 300 and 500 structures). The centroids of each cluster are then used as input to our previously described Variational Bayesian Weighting (VBW) algorithm, which constructs weights for each structure (Fisher *et al.*, 2010, 2012). The output consists of the structures, their weights, the uncertainty parameter (a portion of the Mollack output is shown in Fig. 1), and additional Dirichlet coefficients that can be used to calculate confidence intervals for ensemble average properties. Additionally, ensemble fit statistics are provided, allowing the user to see how well the generated ensemble agrees with the experimental data.

3 Conclusions

We have implemented a web-based tool for the construction of ensembles that model intrinsically disordered proteins. An advantage of the method is that it generates quantitative estimates of the underlying uncertainty in the final model, thereby facilitating rigorous hypothesis testing of observations arising from the model.

Funding

This work was supported by the John Reed Fund, the Paul E. Gray Fund for undergraduate research opportunities (UROP) at MIT, the Lord Foundation UROP Fund, and the National Science Foundation Postdoctoral Research Fellowship in Biology Grant No. 1309247. Additional funding was provided by a Steven G. and Renee Finn Faculty Innovation Award.

Conflict of Interest: none declared.

References

Bashford,D. and Case,D.A. (2000) Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, **51**, 129–152.

Berendsen,H.J.C. *et al.* (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, **91**, 43–56.

Brooks,B.R. *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.

Daughdrill,G.W. *et al.* (2012) Understanding the structural ensembles of a highly extended disordered protein(). *Mol. bioSyst.*, **8**, 308–319.

Fisher,C.K. *et al.* (2010) Modeling intrinsically disordered proteins with Bayesian statistics. *J. Am. Chem. Soc.*, **132**, 14919–14927.

Fisher,C.K. *et al*. (2012) Efficient construction of disordered protein ensembles in a Bayesian framework with optimal selection of conformations. *Pac. Symp. Biocomput.*, 82–93.

Fu,B. *et al*. (2014) MD simulations of intrinsically disordered proteins with replica-averaged chemical shift restraints. *Biophys. J.*, **106**, 481a.

Gurry,T. and Stultz,C.M. (2014) Mechanism of amyloid-beta fibril elongation. *Biochemistry-US*, **53**, 6981–6991.

Jensen,M.R. *et al*. (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J. Am. Chem. Soc.*, **132**, 1270–1272.

Jha,A.K. *et al*. (2005) Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA*, **102**, 13099–13104.

Lazaridis,T. and Karplus,M. (1999) Effective energy function for proteins in solution. *Proteins*, **35**, 133–152.

Marsh,J.A. and Forman-Kay,J.D. (2009) Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J. Mol. Biol.*, **391**, 359–374.

Ozenne,V. *et al*. (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, **28**, 1463–1470.

Tompa,P. and Varadi,M. (2014) Predicting the predictive power of IDP ensembles. *Structure*, **22**, 177–178.

Ulrich,E.L. *et al*. (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.

Varadi,M. *et al*. (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.