BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 3 2013, pages 398–399
doi:10.1093/bioinformatics/bts701

Structural bioinformatics

Advance Access publication December 11, 2012

Characterization of disordered proteins with ENSEMBLE

Mickaël Krzeminski[1], Joseph A. Marsh[1,2,†], Chris Neale[1,2], Wing-Yiu Choy[1,2,‡] and Julie D. Forman-Kay[1,2,*]

[1]Molecular Structure & Function Program, Hospital for Sick Children, Toronto, ON M5G 1X8, Canada and [2]Department of Biochemistry, University of Toronto, Toronto, ON M5S 1A8, Canada

Associate Editor: Anna Tramontano

ABSTRACT

Summary: ENSEMBLE is a computational approach for determining a set of conformations that represents the structural ensemble of a disordered protein based on input experimental data. The disordered protein can be an unfolded or intrinsically disordered state. Here, we introduce the latest version of the program, which has been enhanced to facilitate its general release and includes an intuitive user interface, as well as new approaches to treat data and analyse results.

Availability and implementation: ENSEMBLE is a program implemented in C and embedded in a Perl wrapper. It is supported on main Linux distributions. Source codes and installation files, including a detailed example, can be freely downloaded at http://abragam.med.utoronto.ca/~JFKlab.

Contact: forman@sickkids.ca

Supplementary information: Supplementary Material are available at *Bioinformatics* online.

Received on August 16, 2012; revised on November 6, 2012; accepted on December 1, 2012

1 INTRODUCTION

A detailed understanding of biology requires structural and dynamic information for the proteins in the cell and their dynamic excursions for enzymatic and other functions, as well as their interactions, both transient and long-lived. There has been an explosion of structural and dynamic information on folded protein states that has coincided with the development of computational tools for incorporation of experimental data into algorithms to describe the dominant conformations populated by stable proteins [CNS (Brunger *et al.*, 1998), XPLOR-NIH (Schwieters *et al.*, 2003) and CCP4 (Winn *et al.*, 2011)]. However, unfolded and intrinsically disordered states have not been extensively characterized, although they are highly relevant to both normal and pathological function. Knowledge of unfolded states is important for understanding protein stability, and these states are required to traverse the membrane and are involved in aggregation or degradation in many diseases (Dyson and Wright, 2004). Intrinsically disordered proteins often mediate regulatory protein interactions and can act as entropic springs or steric gates, as well as aggregate in disease (Dunker *et al.*, 2008; Dyson and Wright, 2005; Uversky *et al.*, 2008). Descriptions of the ensembles of highly heterogeneous conformations within disordered states are thus required for correlating dynamic structure with function. A number of algorithms to calculate such ensembles have been developed (Choy and Forman-Kay, 2001; Dedmon *et al.*, 2005; Fisher *et al.*, 2010; Krzeminski *et al.*, 2009; Salmon *et al.*, 2010; Schneider *et al.*, 2012). Our ENSEMBLE approach (Choy and Forman-Kay, 2001; Fisher *et al.*, 2010; Marsh and Forman-Kay, 2009; Marsh and Forman-Kay, 2011; Marsh *et al.*, 2007) was created to facilitate the incorporation of a large number of different types of experimental data into the generation of structural ensembles of disordered protein states.

ENSEMBLE makes use of a switching Monte-Carlo algorithm to choose, from a large set of conformations, an ensemble for which the *in silico* back-calculated data fit all available experimental data. Since its first version (Choy and Forman-Kay, 2001), we have developed a more robust algorithm, enhanced the sampling of conformational space and expanded the types of experimental data used. Applications of ENSEMBLE to the unfolded state of the drkN SH3 domain (Choy and Forman-Kay, 2001; Marsh and Forman-Kay, 2009; Marsh *et al.*, 2007) and the intrinsically disordered states of the cell cycle protein Sic1 (Mittag *et al.*, 2010) and various protein phosphatase 1 regulators (Marsh *et al.*, 2010; Pinheiro *et al.*, 2011) have also been described. Importantly, using cross-validation and synthetic experimental restraints calculated from simulated ensembles, we demonstrated that ENSEMBLE is capable of accurate modeling of the secondary structure, molecular size distribution and tertiary contacts of disordered proteins, if sufficient data is incorporated; tertiary structure, in particular, is highly dependent on the number of distance restraints used (Marsh and Forman-Kay, 2012). Most recently, we have significantly improved the user interface and the stability of the program. Thus, we are now making the first official distribution (version 2.1) available online at http://abragam.med.utoronto.ca/~JFKlab.

2 GENERAL SCHEME

ENSEMBLE currently accommodates ten different types of data (*modules*), primarily from nuclear magnetic resonance experiments [chemical shift, residual dipolar coupling, J-coupling, R_2 relaxation rate, paramagnetic relaxation enhancement, paramagnetic relaxation enhancement ratio, nuclear Overhauser effect, solvent accessibility and hydrodynamic radius (R_h)] and small

*To whom correspondence should be addressed.
†Present address: MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK.
‡Present address: Department of Biochemistry, University of Western Ontario, London, ON N6A 5C1, Canada.

angle X-ray scattering. R_h can also be obtained from size exclusion chromatography or dynamic light scattering. The expected observables for each type of data are back-calculated for each conformer of a large set of input structures within the specific module for this data type, in some cases using external programs ShiftX (Neal *et al.*, 2003), HYDROPRO (Ortega *et al.*, 2011) and CRYSOL (Svergun *et al.*, 1995).

The algorithm that governs ENSEMBLE is more efficient when searching for best-fitting structures within a set containing <5000 conformations. Obtaining a reasonable sampling of the theoretical conformational space for disordered proteins, however, is not possible with such a low number of structures. ENSEMBLE may be used for a variety of protein conformational states including folded states—also inherently dynamic, in which case this pool may be large enough. This led us to construct a Perl wrapper, which is split into parts (see Supplementary Material).

Conformer management: A large set of structures, called the initial soup, is populated by structures provided by the user (e.g. structures from simulations) and/or regularly replenished with conformers generated by TraDES (Feldman and Hogue, 2000; Feldman and Hogue, 2002). We found that at least 100 000 structures in the initial soup is sufficient for conformational sampling. The C core of ENSEMBLE acts on a subset of the initial soup conformations, called the initial pool, containing randomly selected structures from the initial soup. Each conformer can be selected multiple times, but the algorithm favors conformers that have been selected fewer times. The initial pool is prevented from becoming larger than 5000 conformers by periodic removal of those that do not contribute to fitting the experimental data. To increase sampling of conformational space, selected structures of the currently best fitting ensemble can be slightly randomized with TraDES and added to the initial soup and initial pool.

The selection process: The C core of ENSEMBLE performs the selection of an ensemble from all conformers present in the initial pool. Each module is assigned a weight that determines its importance in the scoring function. A module fits the experimental data when its score is lower than a threshold value automatically determined by ENSEMBLE, reflecting the back-calculated data for the selected conformers matching the experimental data. The module weights are changed and ENSEMBLE is run again until all modules fit the experimental data. The user can also incorporate restraints incrementally.

The final ensemble analysis: Once the selected ensemble fits all the experimental data, this final ensemble is analyzed. ENSEMBLE includes a set of scripts for analysis (C_α–C_α distances, radius of gyration, secondary structure distribution) and management of binary structure and data files generated by the program. These will be expanded in the future.

The release of this version of ENSEMBLE provides tools that can be broadly used to enhance the understanding of structural and dynamic properties of disordered proteins and their correlations with biological function and disease.

Funding: This work was funded by grants to J.D.F.-K. from the Natural Sciences and Engineering Research Council, Canadian Institutes of Health Research and Cystic Fibrosis Canada.

Conflict of interest: None declared.

REFERENCES

Brunger,A.T. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta. Crystallogr. D Biol. Crystallogr.*, **54**, 905–921.

Choy,W.Y. and Forman-Kay,J.D. (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.*, **308**, 1011–1032.

Dedmon,M.M. *et al.* (2005) Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.*, **127**, 476–477.

Dunker,A.K. *et al.* (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.

Dyson,H.J. and Wright,P.E. (2004) Unfolded proteins and protein folding studied by NMR. *Chem. Rev.*, **104**, 3607–3622.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.*, **6**, 197–208.

Feldman,H.J. and Hogue,C.W. (2000) A fast method to sample real protein conformational space. *Proteins*, **39**, 112–131.

Feldman,H.J. and Hogue,C.W. (2002) Probabilistic sampling of protein conformations: new hope for brute force? *Proteins*, **46**, 8–23.

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

Krzeminski,M. *et al.* (2009) MINOES: a new approach to select a representative ensemble of structures in NMR studies of (partially) unfolded states. *Application to Delta25-PYP. Proteins*, **74**, 895–904.

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.

Marsh,J.A. and Forman-Kay,J.D. (2012) Ensemble modeling of protein disordered states: experimental restraint contributions and validation. *Proteins*, **80**, 556–572.

Marsh,J.A. *et al.* (2010) Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure*, **18**, 1094–1103.

Marsh,J.A. *et al.* (2007) Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *J. Mol. Biol.*, **367**, 1494–1510.

Mittag,T. *et al.* (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure*, **18**, 494–506.

Neal,S. *et al.* (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, **26**, 215–240.

Ortega,A. *et al.* (2011) Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.*, **101**, 892–898.

Pinheiro,A.S. *et al.* (2011) Structural signature of the MYPT1-PP1 interaction. *J. Am. Chem. Soc.*, **133**, 73–80.

Salmon,L. *et al.* (2010) NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.*, **132**, 8407–8418.

Schneider,R. *et al.* (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. Biosyst.*, **8**, 58–68.

Schwieters,C.D. *et al.* (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson*, **160**, 65–73.

Svergun,D. *et al.* (1995) CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J.Appl. Crystallogr.*, **28**, 768–773.

Uversky,V.N. *et al.* (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, **37**, 215–246.

Winn,M.D. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta. Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.