Title

Local Disordered Region Sampling (LDRS) for ensemble modeling of proteins with experimentally undetermined or low confidence prediction segments.

Journal

Bioinformatics, 39(12)

Authors

Liu, Zi

Teixeira, João

Zhang, Oufan

et al.

Publication Date

2023-12-01

DOI

10.1093/bioinformatics/btad739

Peer reviewed

Bioinformatics, 2023, **39(12)**, btad739
https://doi.org/10.1093/bioinformatics/btad739
Advance Access Publication Date: 7 December 2023
Applications Note

OXFORD

Structural bioinformatics

Local Disordered Region Sampling (LDRS) for ensemble modeling of proteins with experimentally undetermined or low confidence prediction segments

Zi Hao Liu [1,2], João M.C. Teixeira[1], Oufan Zhang[3,4], Thomas E. Tsangaris [5,6], Jie Li[3,4], Claudiu C. Gradinaru[5,6], Teresa Head-Gordon [3,4,7,8], Julie D. Forman-Kay[1,2,*]

[1]Molecular Medicine Program, Hospital for Sick Children, Toronto, ON M5G 0A4, Canada
[2]Department of Biochemistry, University of Toronto, Toronto, ON M5S 1A8, Canada
[3]Pitzer Center for Theoretical Chemistry, University of California, Berkeley, Berkeley, CA 94720, United States
[4]Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720-1460, United States
[5]Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada
[6]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada
[7]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA 94720-1462, United States
[8]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720-1762, United States

*Corresponding author. Molecular Medicine Program, The Hospital for Sick Children, 686 Bay St., Toronto, ON M5G 0A4, Canada. E-mail: forman@sickkids.ca

Associate Editor: Lenore Cowen

Abstract

Summary: The Local Disordered Region Sampling (LDRS, pronounced *loaders*) tool is a new module developed for IDPConformerGenerator, a previously validated approach to model intrinsically disordered proteins (IDPs). The IDPConformerGenerator LDRS module provides a method for generating all-atom conformations of intrinsically disordered protein regions at N- and C-termini of and in loops or linkers between folded regions of an existing protein structure. These disordered elements often lead to missing coordinates in experimental structures or low confidence in predicted structures. Requiring only a pre-existing PDB or mmCIF formatted structural template of the protein with missing coordinates or with predicted confidence scores and its full-length primary sequence, LDRS will automatically generate physically meaningful conformational ensembles of the missing flexible regions to complete the full-length protein. The capabilities of the LDRS tool of IDPConformerGenerator include modeling phosphorylation sites using enhanced Monte Carlo-Side Chain Entropy, transmembrane proteins within an all-atom bilayer, and multi-chain complexes. The modeling capacity of LDRS capitalizes on the modularity, the ability to be used as a library and via command-line, and the computational speed of the IDPConformerGenerator platform.

Availability and implementation: The LDRS module is part of the IDPConformerGenerator modeling suite, which can be downloaded from GitHub at https://github.com/julie-forman-kay-lab/IDPConformerGenerator. IDPConformerGenerator is written in Python3 and works on Linux, Microsoft Windows, and Mac OS versions that support DSSP. Users can utilize LDRS's Python API for scripting the same way they can use any part of IDPConformerGenerator's API, by importing functions from the "idpconfgen.ldrs_helper" library. Otherwise, LDRS can be used as a command line interface application within IDPConformerGenerator. Full documentation is available within the command-line interface as well as on IDPConformerGenerator's official documentation pages (https://idpconformergenerator.readthedocs.io/en/latest/).

Atomistic structures of proteins, which are computational models of real proteins, have been a central goal of the field of structural biology to generate hypotheses and deepen our understanding of the relationship between protein structure and function. Experimental structures determined by X-ray crystallography, NMR spectroscopy and cryo-electron microscopy (cryoEM) have provided incredible structural insights (Dokholyan 2020, Burley *et al.* 2022). Most recently, accurate protein structure predictions have become available using machine learning methods such as AlphaFold (Jumper *et al.* 2021), RoseTTAFold (Baek *et al.* 2021), and ESMFold (Lin *et al.* 2023).

Intrinsically disordered protein regions (IDRs) are not visible by X-ray crystallography and cryoEM because data averaging due to conformational heterogeneity leads to missing electron density and, hence, missing coordinates in the final models (Villarreal and Stewart 2014, Djinovic-Carugo and Carugo 2015, Nwanochie and Uversky 2019). Computational predicted structures contain coordinates for IDRs, but these coordinates generally have low confidence predictions and are not likely representative (Ruff and Pappu 2021). Since about 30% of residues within the human proteome are expected to be within IDRs (Tsang *et al.* 2020), obtaining structural insights from ensemble models of these flexible regions has become a focus.

A variety of computational sampling methods are available to model single-chain intrinsically disordered proteins (IDPs) ensembles, including IDPConformerGenerator (Teixeira *et al.* 2022), Flexible-meccano (Ozenne *et al.* 2012), TraDES (Feldman and Hogue 2000, 2002), and others (Shrestha *et al.* 2021, Karamanos *et al.* 2022). FastFloppyTail (Ferrie and

Received: 25 July 2023; **Revised:** 30 October 2023; **Editorial Decision:** 29 November 2023; **Accepted:** 6 December 2023

Petersson 2020) can be used to generate IDPs or IDRs at the N- and C-termini of a folded domain. Currently, though, there is no easy-to-use modular method of modeling conformational ensembles of (i) IDRs between two folded domains within an experimental structure, (ii) IDRs representing low confidence regions of predicted structures, (iii) IDRs within transmembrane proteins, or (iv) IDRs found in dynamic protein complexes. While cyclic coordinate descent (CCD) and kinematic closure methods (KIC) (Canutescu and Dunbrack 2003, Boomsma and Hamelryck 2005, Stein and Kortemme 2013, O'Donnell *et al.* 2022) are able to model all-atom missing protein regions (i.e. breaks in the protein chain), these approaches are limited to 12 residues. In addition, existing "loop" closure methods such as CCD do not rely on statistical relationships between backbone torsion angles and protein sequence, resulting in the lack of sampling of realistic fractional secondary structure.

Here, we present the Local Disordered Region Sampling (LDRS) module of IDPConformerGenerator, available in the IDPConformerGenerator v0.7.10 update. The LDRS module enables users to model all-atom sidechain-inclusive ensembles of N-terminal and C-terminal IDRs (N-IDRs and C-IDRs, respectively), as well as IDRs between folded elements (L-IDRs, for loops or linkers). Because LDRS is a module of IDPConformerGenerator, it uses the knowledge-based modeling engine of IDPConformerGenerator, which generates physically meaningful ensembles based on the torsion angle sampling driven by the statistical relationship between protein sequence and backbone torsion angles along with other chain geometry within the RCSB PDB (Burley *et al.* 2022). While the resulting IDR conformer ensembles have not been refined with experimental data and are thus not necessarily accurate, they are expected to be representative, based on our previously reported observed agreement between IDPConformerGenerator ensemble structural properties and experiment (Teixeira *et al.* 2022). LDRS has been developed as a modular command-line interface and Python library within the IDPConformerGenerator platform (Teixeira *et al.* 2022), exploiting the flexibility and modularity of IDPConformerGenerator's design. The initial release of IDPConformerGenerator could only model isolated IDPs and IDRs, but with the LDRS update, using the new "ldrs" subclient, users are able to model IDRs within the context of proteins having folded domains. Furthermore, these IDRs can be modeled automatically in the context of a lipid bilayer surrounding a membrane-embedded domain, as well as in multi-chain dynamic complex systems. As of v0.7.10, both PDB and PDBx/mmCIF formatted initial structural templates with missing atomic coordinates for the IDRs to be modeled are accepted.

To showcase the multiple applications of LDRS, we have modeled five protein systems having different combinations of N-IDR, L-IDR, and C-IDR cases. We use structures from the RCSB PDB (Berman *et al.* 2000) having missing coordinates due to lack of data and from the AlphaFold structure prediction database (Jumper *et al.* 2021, Varadi *et al.* 2022), from which we removed the low confidence residues. We modeled missing residues as ensembles using LDRS. Sidechain atoms were modeled within IDPConformerGenerator using Monte Carlo-Side Chain Entropy (MC-SCE) (Bhowmick and Head-Gordon 2015), and we exploited recent enhancements to MC-SCE that enable modeling post-translational modifications (PTMs), including phosphorylation, methylation, N6-carboxylysine, and hydroxylation (Han and Martinage 1992, Sirota *et al.* 2015). We also demonstrate the capabilities to model IDRs in transmembrane proteins in the context of a phospholipid bilayer and within multi-chain dynamic complexes. Full-length protein ensembles of these systems have not been previously modeled/deposited on the Protein Ensemble Database (PED) (Lazar *et al.* 2021, Ghafouri *et al.* 2023) at the time of writing. We deposited all modeled ensembles for this project and have listed the PED IDs in Fig. 1 and Supplementary Fig. S2. The primary sequence schematics for the folded and IDR elements for the studied systems are shown in Supplementary Fig. S1.

Figure 1A presents a combination of the N-IDR and C-IDR cases occurring in the 5-fold phosphorylated eukaryotic translation initiation factor 4E-binding protein 2 (5p 4E-BP2), based on the NMR structure (PDB ID 2MX4) (Bah *et al.* 2015). 4E-BP2 is largely disordered but has a conditionally folded ~40-residue domain upon phosphorylation (Bah *et al.* 2015, Dawson *et al.* 2020), leading to an N-IDR length of 18 residues and a C-IDR length of 59 residues (including 3 phosphorylation sites) surrounding the folded domain (containing 2 phosphorylation sites). The all-atom coordinates generated for 5p 4E-BP2 include the five phosphate groups. For the single L-IDR case (Fig. 1B), we have modeled the STAS (sulfate transporter anti-sigma) domain of SLC26A9 (solute carrier family 26 member 9) including 86 residues not found in the electron density of the X-ray structure (PDB ID 7CH1) (Chi *et al.* 2020).

Figure 1C presents the transmembrane α_{2A} adrenergic receptor predicted by AlphaFold (entry P08913), which has extensive regions with low confidence residues (pLDDT < 70) (Jumper *et al.* 2021, Varadi *et al.* 2022). We removed coordinates for these low confidence residues using a simple script that is easily modified to apply to other predicted structures (see Supplementary Information), leading to a 40-residue N-IDR, two L-IDRs of 18 residues and 125 residues, and a 9-residue C-IDR to be modeled by LDRS. Initially, the lipid bilayer was added using OPM (Orientations of Proteins in Membranes) (Lomize *et al.* 2012) and the CHARMM-GUI (Jo *et al.* 2008). The atomic coordinates of the lipid bilayer are then accounted for in LDRS for steric clash checking purposes. While only backbones were modeled for the IDRs, all atomic coordinates of the lipids are present in each of the LDRS models of the α_{2A} adrenergic receptor.

Figure 1D shows an ensemble model of the dynamic interaction between the eukaryotic translation initiation factor 4E (eIF4E) and nonphosphorylated 4E-BP2 (Lukhele *et al.* 2013), with 4E-BP2 represented as having two fixed binding sites to eIF4E highlighted in dark grey. The backbone atoms for the N-IDR (54 residues), L-IDR (20 residues), and C-IDR (40 residues) of nonphosphorylated 4E-BP2 and for the N-IDR (35 residues) and L-IDR (9 residues) of eIF4E have been modeled using LDRS. The template for the complex is derived from an X-ray crystal structure of 4E-BP1 bound to eIF4E (PDB ID 4UED) (Peter *et al.* 2015), removing residues of 4E-BP2 that are extremely broadened in NMR spectra of the complex (Lukhele *et al.* 2013). Lastly, we have modeled the cellular tumor antigen p53 (p53) with all-atom conformations using the full-length predicted structure from AlphaFold (entry P04637) but removing low confidence regions (Supplementary Fig. S2). Note that the α_{2A} adrenergic receptor transmembrane protein and the 4E-BP2:eIF4E complex backbone ensembles can have all-atom conformers by adding a sidechain packing step akin to the other presented systems, as IDPConformerGenerator LDRS can

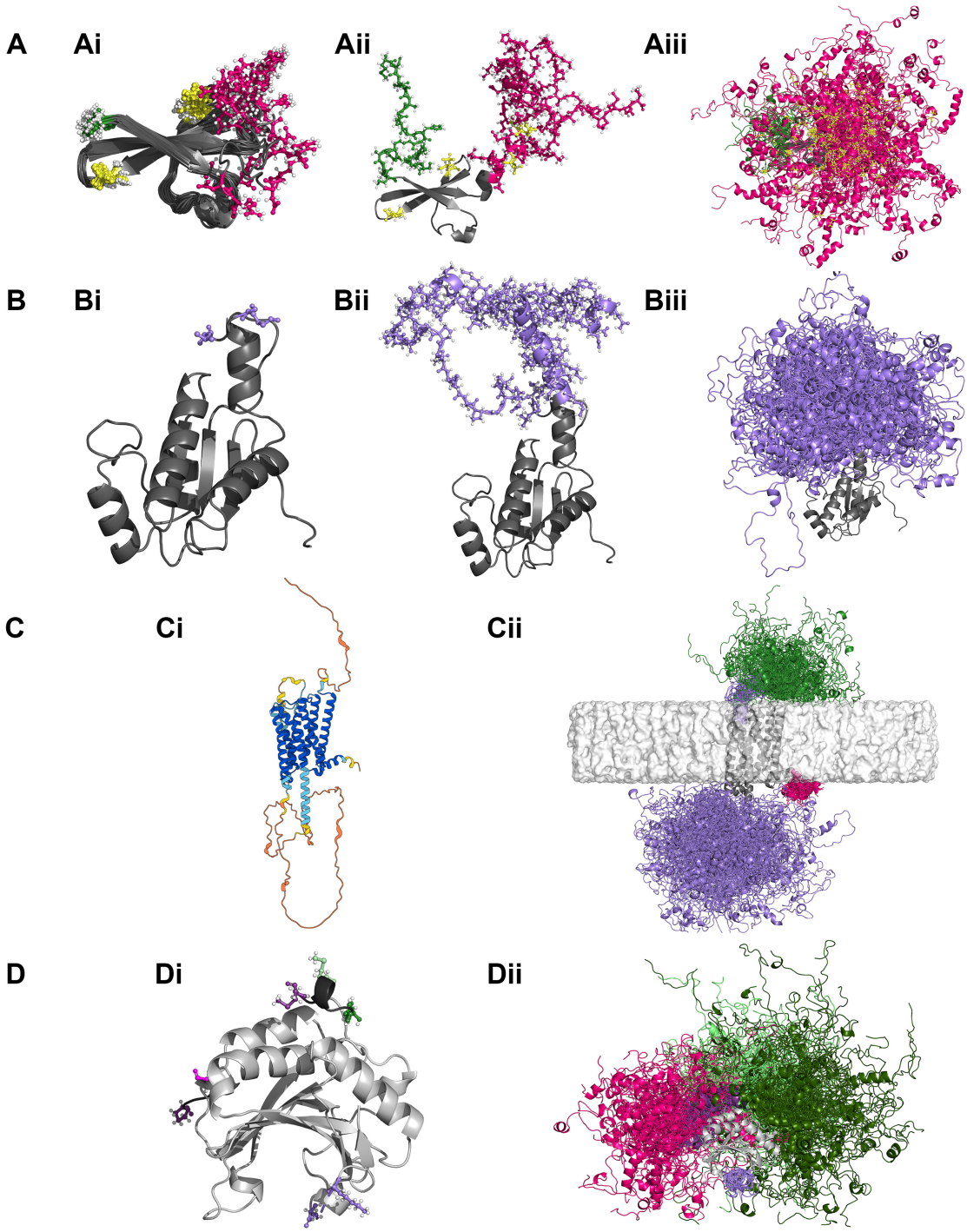

Figure 1. Diverse full-length all-atom and backbone-only ensembles generated using LDRS. Models are drawn as ribbons; disordered regions have additional ball-and-stick models. Pre-existing solved or predicted structure is shown in grey, N-IDRs in green, L-IDRs in purple, and C-IDRs in magenta. Hydrogen atoms are white in the ball-and-stick models. (A) Five-fold phosphorylated 4E-BP2 (5p 4E-BP2, PED ID PED00436): (Ai) NMR structure (20 poses from PDB ID 2MX4) with ball-and-stick representations of the terminal residues, N-terminal proline in green and C-terminal arginine in magenta. Phosphorylated residues within the folded domain (Thr35, Thr46) are shown in yellow. (Aii) A single full-length all-atom conformer with ball-and-stick representations of the IDRs including phosphate residues (Ser65, Thr70, Ser83, in yellow) modeled using LDRS together with MC-SCE. (Aiii) Ensemble of 100 (of 1828 calculated) all-atom conformers of 5p 4E-BP2 calculated with LDRS. (B) SLC26A9 STAS domain (PED ID PED00437): (Bi) X-ray structure (PDB ID 7CH1) with ball-and-stick representations of the residues immediately surrounding the missing loop colored for placement of the missing L-IDR. (Bii) A single all-atom model of the full-length domain with ball-and-stick representation of the LDRS-generated L-IDR missing in panel Bi. (Biii) Ensemble of 100 (of 692 calculated) all-atom conformers. (C) α_{2A} adrenergic receptor (PED ID PED00435): (Ci) Structure predicted by AlphaFold (entry P08913), colored according to AlphaFold confidence, with lower confidence regions (pLDDT < 70) in orange and yellow (Jumper *et al.* 2021, Varadi *et al.* 2022). (Cii) Ensemble of 100 backbone structures (of 2000 calculated) with AlphaFold lower confidence regions (pLDDT < 70) modeled by LDRS in the context of a bilayer (in light grey). (D) 4E-BP2: eIF4E complex (PED ID PED00434): (Di) Homology model based on the 4E-BP1: eIF4E X-ray structure (PDB ID 4UED) with 4E-BP2 elements to be modeled as fixed in dark grey and eIF4E in light grey (see Supplementary Information for more details). Ball-and-stick representations of the terminal residues of IDRs to be modeled for eIF4E in light colors and for 4E-BP2 in dark. (Dii) Ensemble of 100 (of 2000 calculated) full-length backbone conformers of the 4E-BP2: eIF4E complex with IDRs modeled by LDRS. eIF4E N-IDRs shown in light green and L-IDRs in light purple. 4E-BP2 N-IDRs shown in dark green, L-IDRs in dark purple, and C-IDRs in magenta.

automatically detect lipid bilayer boundaries and build on templates with multiple folded chains.

The fraction secondary structure and Ramachandran plots for the calculated models of these proteins showcase the extent of sampling (Supplementary Figs S3–S5). A strength of LDRS is its ability to sample transient helical secondary structure at the boundary of fragment attachment, based on the sequence-based torsion angle builder of IDPConformerGenerator. For example, we observe fractional α-helix secondary structures extending from the existing chains of 5p 4E-BP2 and the SLC26A9 STAS domain (Fig. 1Aii and Bii). This extension of helical secondary structure is captured by DSSP (Kabsch and Sander 1983) as seen in Supplementary Fig. S3. Relative speeds of conformer generation for each system demonstrate the efficiency of the modeling, but are dependent on IDR length and the restrictions from the folded domain and lipid bilayer (Supplementary Table S1).

For ensemble models of IDRs of AlphaFold-predicted structures (Jumper *et al.* 2021, Varadi *et al.* 2022), we find that the conformers exhibit a range of secondary structures (Supplementary Figs S3.1 and S3.2), as expected based on IDPConformerGenerator's sampling of PDB-derived torsion angles, compared to the primarily coil or loop structures of the initial AlphaFold model (Supplementary Fig. S3.3). In addition, Cα-Cα distances between modeled IDRs and the folded template sample a much broader range than the initial AlphaFold model, including in some cases instances with closer contacts reflective of potential interactions sampled by IDPConformerGenerator (Supplementary Figs S6.1 and S6.2).

A graphical summary of the LDRS method is shown in Supplementary Figs S7.1 and S7.2. The core of the LDRS approach is based on the Kabsch algorithm (Kabsch 1976), a mathematical protocol that ensures the alignment and connectivity of protein backbone fragments. First, LDRS identifies missing residues by comparing the given input PDB structure and full-length sequence. LDRS then builds the missing residues as isolated single chains using the IDPConformerGenerator (Teixeira *et al.* 2022) builder. Single chains are oriented to the sites of N-IDR, L-IDR, and C-IDR cases using the Kabsch algorithm, and a flexible van der Waals radii (Tsai *et al.* 1999) clash check is employed to remove chains that clash with the input structure. Although sidechains can be built during the initial modeling step, we recommend modeling the backbones first before the downstream sidechain and post-translational modification modeling with MC-SCE (Bhowmick and Head-Gordon 2015) to eliminate sidechain clashes. Phosphorylated Ser and Thr sidechains are added after sampling torsion angles related to unmodified Ser and Thr residues, as well as pSer and pThr residues, from the RCSB PDB (Berman *et al.* 2000, Burley *et al.* 2022), due to the lack of data for backbone specific torsion angles of phosphorylated residues.

To model IDRs between folded elements (L-IDR case), we have developed the *next seeker* algorithm in LDRS (see Supplementary Fig. S7.1). After generating ensembles of full-length fragments of these IDRs representing missing residues from both sides of the gap, *next seeker* identifies fragment pairs that can close the chain. This is done by verifying that the Cα(i)-C(i)-N(i + 1) bond angle, ω backbone torsion angle, and bond lengths ($d_{C\text{-}N}$, $d_{C\text{-}C\alpha}$) of residues i and i + 1 from each of the respective pairs comply with average values observed in the IDPConformerGenerator database. This database was generated from nonredundant PDB IDs of X-ray crystal structures with resolutions better than or equal to 1.8 Å. After a match has been found, *next seeker* can remodel the carbonyl oxygen and the amino hydrogen at the point of closure.

As LDRS appends a new library of functions into the IDPConformerGenerator API, this generalized tool can be used to model intricate protein systems limited only by the imagination of the user. For example, using the scripts provided in the Supplementary Material archive as a starting point, LDRS could be used to model ensembles of transmembrane protein interactions comprising several different protein chains.

We envision that the LDRS module within IDPConformer Generator will be a useful tool to generate ensemble representations of highly flexible loops and tails that are poorly represented by data or prediction methods. The agreement of *ab initio* IDPConformerGenerator ensembles with experimental data (Teixeira *et al.* 2022) suggests that the IDRs built with LDRS should be representative. If experimental data on torsion angle preferences (i.e., NMR chemical shifts and J-coupling data) are available for the IDRs, the LDRS module in conjunction with the CSSS (custom secondary structure sampling) module (Teixeira *et al.* 2022) within IDPConformerGenerator can build IDRs with these preferences. Furthermore, all-atom conformer ensembles generated with LDRS can be input into sub-setting or reweighting protocols using experimental data, including nuclear magnetic resonance (NMR) spectroscopy, small-angle X-ray scattering (SAXS), and fluorescence resonance energy transfer (FRET) (Bottaro *et al.* 2020, Gomes *et al.* 2020, Lincoff *et al.* 2020, Liu *et al.* 2023, Tsangaris *et al.* 2023) to generate more realistic ensemble representations that agree with these data. Future development of the LDRS toolkit within the IDPConformerGenerator platform will include streamlining generation of dynamic complexes involving IDPs and IDRs, improving the efficiency of sidechain packing by MC-SCE (Bhowmick and Head-Gordon 2015) and enhancing its capabilities to represent additional post-translational modification types. The expanding toolkit of IDPConformerGenerator will facilitate structural modeling of the many IDRs present in human and other proteomes, providing valuable insights into structure-dynamics-disorder-function relationships.

Acknowledgements

We acknowledge Philip Rößler for helping identify the α2A adrenergic receptor system as a transmembrane protein case, and Zhi Wei Zeng for assistance with the protocol to model the bilayer seen in the α2A adrenergic receptor system.

Supplementary data

Supplementary data are available at *Bioinformatics* online, including all the conformational ensembles and analyses described in this paper.

Conflict of interest

None declared.

Funding

This work was supported by the National Institutes of Health [5R01GM127627-04 and 2R01GM127627-05 to T.H.-G. and J.D.F.-K.]; the Natural Sciences and Engineering Research Council of Canada [2016-06718 to J.D.F.-K., and

RGPIN-2023-04864 to C.C.G.]; and the Canada Research Chairs Program [to J.D.F.-K.].

References

Baek M, DiMaio F, Anishchenko I *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.

Bah A, Vernon RM, Siddiqui Z *et al.* Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature* 2015;**519**:106–9.

Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.

Bhowmick A, Head-Gordon T. A Monte Carlo method for generating side chain structural ensembles. *Structure* 2015;**23**:44–55.

Boomsma W, Hamelryck T. Full cyclic coordinate descent: solving the protein loop closure problem in Cα space. *BMC Bioinformatics* 2005;**6**:159.

Bottaro S, Bengtsen T, Lindorff-Larsen K. Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. *Methods Mol Biol* 2020;**2112**:219–40.

Burley SK, Berman HM, Duarte JM *et al.* Protein data bank: a comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules* 2022;**12**:1425.

Canutescu AA, Dunbrack RL. Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci* 2003;**12**:963–72.

Chi X, Jin X, Chen Y *et al.* Structural insights into the gating mechanism of human SLC26A9 mediated by its C-terminal sequence. *Cell Discov* 2020;**6**:55.

Dawson JE, Bah A, Zhang Z *et al.* Non-cooperative 4E-BP2 folding with exchange between eIF4E-binding and binding-incompatible states tunes cap-dependent translation inhibition. *Nat Commun* 2020;**11**:3146.

Djinovic-Carugo K, Carugo O. Missing strings of residues in protein crystal structures. *Intrinsically Disord Proteins* 2015;**3**:e1095697.

Dokholyan NV. Experimentally-driven protein structure modeling. *J Proteomics* 2020;**220**:103777.

Feldman H, Hogue C. A fast method to sample real protein conformational space. *Proteins* 2000;**39**:112–31.

Feldman HJ, Hogue CWV. Probabilistic sampling of protein conformations: new hope for brute force? *Proteins* 2002;**46**:8–23.

Ferrie JJ, Petersson EJ. A unified de novo approach for predicting the structures of ordered and disordered proteins. *J Phys Chem B* 2020;**124**:5538–48.

Ghafouri H, Lazar T, Del Conte A *et al.*; PED Consortium. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res* 2023;gkad947.

Gomes G-NW, Krzeminski M, Namini A *et al.* Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc* 2020;**142**:15697–710.

Jo S, Kim T, Iyer VG *et al.* CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem* 2008;**29**:1859–65.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst A* 1976;**32**:922–3.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**:2577–637.

Karamanos TK, Kalverda AP, Radford SE. Generating ensembles of dynamic misfolding proteins. *Front Neurosci* 2022;**16**:881534.

Han KK, Martinage A. Post-translational chemical modification(S) of proteins. *Int J Biochem* 1992;**24**:19–28.

Lazar T, Martínez-Pérez E, Quaglia F *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res* 2021;**49**:D404–11.

Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30.

Lincoff J, Haghighatlari M, Krzeminski M *et al.* Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun Chem* 2020;**3**:1–12.

Liu ZH, Zhang O, Teixeira J C *et al.* SPyCi-PDB: a modular command-line interface for back-calculating experimental datatypes of protein structures. *JOSS* 2023;**8**:4861.

Lomize MA, Pogozheva ID, Joo H *et al.* OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* 2012;**40**:D370–6.

Lukhele S, Bah A, Lin H *et al.* Interaction of the eukaryotic initiation factor 4E with 4E-BP2 at a dynamic bipartite interface. *Structure* 2013;**21**:2186–96.

Nwanochie E, Uversky VN. Structure determination by single-particle cryo-electron microscopy: only the sky (and intrinsic disorder) is the limit. *Int J Mol Sci* 2019;**20**:4186.

O'Donnell T, Robert CH, Cazals F. Tripeptide loop closure: a detailed study of reconstructions based on Ramachandran distributions. *Proteins Struct Funct Bioinf* 2022;**90**:858–68.

Ozenne V, Bauer F, Salmon L *et al.* Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 2012;**28**:1463–70.

Peter D, Igreja C, Weber R *et al.* Molecular architecture of 4E-BP translational inhibitors bound to eIF4E. *Mol Cell* 2015;**57**:1074–87.

Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol* 2021;**433**:167208.

Shrestha UR, Smith JC, Petridis L. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun Biol* 2021;**4**:243–8.

Sirota FL, Maurer-Stroh S, Eisenhaber B *et al.* Single-residue posttranslational modification sites at the N-terminus, C-terminus or in-between: to be or not to be exposed for enzyme access. *Proteomics* 2015;**15**:2525–46.

Stein A, Kortemme T. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One* 2013;**8**:e63090.

Teixeira JMC, Liu ZH, Namini A *et al.* IDPConformerGenerator: a flexible software suite for sampling the conformational space of disordered protein states. *J Phys Chem A* 2022;**126**:5985–6003.

Tsai J, Taylor R, Chothia C *et al.* The packing density in proteins: standard radii and volumes11 Edited by J. M. Thornton. *J Mol Biol* 1999;**290**:253–66.

Tsang B, Pritišanac I, Scherer SW *et al.* Phase separation as a missing mechanism for interpretation of disease mutations. *Cell* 2020;**183**:1742–56.

Tsangaris TE, Smyth S, Gomes G-NW *et al.* Delineating structural propensities of the 4E-BP2 protein via integrative modeling and clustering. *J Phys Chem B* 2023;**127**:7472–86.

Varadi M, Anyango S, Deshpande M *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.

Villarreal SA, Stewart PL. CryoEM and image sorting for flexible protein/DNA complexes. *J Struct Biol* 2014;**187**:76–83.