**Structural bioinformatics**

**ProDy 2.0: Increased Scale and Scope after 10 Years of Protein Dynamics Modelling with Python**

She Zhang1,*,2†, James M. Krieger1,*,3†, Yan Zhang1, Cihan Kaya1,4, Burak Kaynak1, Karolina Mikuliska-Ruminska1,5, Petra Doruker1, Hongchun Li1,6†, Ivet Bahar1†

1Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
1To whom correspondence should be addressed.
*These authors contributed equally to this work.
†To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

Abstract

Summary: ProDy, an integrated API developed for modelling and analysing protein dynamics, has significantly evolved in recent years in response to the growing data and needs of computational biology community. We present major developments that led to ProDy 2.0: (i) improved interfacing with databases and parsing new file formats, (ii) SignDy for signature dynamics of protein families, (iii) CryoDy for collective dynamics of supramolecular systems using cryo-EM density maps, and (vi) essential site scanning analysis (ESSA) for identifying sites essential to modulating global dynamics.

Availability and Implementation: ProDy is open-source and freely available under MIT License from [https://github.com/prody/ProDy](https://github.com/prody/ProDy). Contact: [bahar@pitt.edu](mailto:bahar@pitt.edu), [shezhang620@gmail.com](mailto:shezhang620@gmail.com), [jamesmkrieger@gmail.com](mailto:jamesmkrieger@gmail.com), [hongchun.li@siat.ac.cn](mailto:hongchun.li@siat.ac.cn)

Supplementary information: Supplementary data are available at Bioinformatics online, and tutorials at [http://prody.csb.pitt.edu/](http://prody.csb.pitt.edu/).

Introduction

Proteins are dynamic entities. Their structural dynamics is essential to their myriad functions (Bahar et al., 2017). The ProDy application programming interface (API) in Python was introduced in 2011 to provide a unified environment for analyses of protein dynamics and mechanisms which lay the framework for their biological activities (Bakan et al., 2011). The API was upgraded in 2014 by adding a new module, Evol, to enable sequence evolutionary analysis complementing that of structural dynamics (Bakan et al., 2014). The original API featured functions and data structures for spectral mode decomposition and/or normal mode analysis (NMA) based on elastic network models (ENMs, including the Anisotropic Network Model (ANM) (Attilgan et al., 2001) and Gaussian Network Model (GNM) (Bahar et al., 1997)), and principal component analysis (PCA) of experimental structures, allowing users to evaluate and visualize structural dynamics, and make rigorous comparisons of motions derived from experiments and computations. The API has been significantly upgraded since then, and has found wide utility, evidenced by more than 2 million downloads from PyPI and 150,000+ unique website visits.

The current Application Note aims at providing a summary of recent updates. We focus here on three recent modules implemented in ProDy: evaluation of the signature dynamics of protein families (SignDy) (Zhang et al., 2019); characterization of the collective dynamics of supramolecular structures resolved by cryo-EM, using electron density maps as inputs to construct ENMs (Zhang et al., 2020), and essential site scanning analysis (ESSA) (Kaynak et al., 2020); along with general upgrades in ProDy core architecture, yielding a new generation of ProDy 2.0.

1. Inputs and Outputs: new file parsers and interfaces

The traditional input for ProDy is a PDB file, either provided by the user or retrieved from the Protein Data Bank (PDB) using an ID or sequence, and the output is structural dynamics. The outputs are various objects, relating to coordinates, sequences and alignments, ensembles, and normal modes (Supplementary Figure 1), as well as various plots facilitated by integration with numeric and scientific Python libraries, NumPy (Harris et al., 2020) and SciPy (Virtanen et al., 2020), and plotting library Matplotlib (Hunter, 2007), and the visualisation tool NMWiz as a VMD plug-in (Bakan et al., 2014). Data-handling capabilities of ProDy have been significantly enhanced in version 2.0. For example, development of family-based analysis (in SignDy) led to integration with diverse databases and servers, enabling users to find similar structures upon inputting a single sequence or ID and calculate functional properties for the entire protein family. Other interfaces added include UniProt and QuartataWeb (Li et al., 2020) for drug-target interactions. New parsers include those for the PDBx/mmCIF format (Adams et al., 2019) and cryo-EM maps in MRC2014 format (Cheng et al., 2015) from the EMDataBank.

1 Present address: 1OpenEye Scientific Software Inc., Santa Fe, NM, USA; 2Pfizer, Groton, CT, USA; 3Institute of Physics, Faculty of Physics Astronomy and Informatics, Nicolaus Copernicus University in Torun, Poland; 4Research Center for Computer-Aided Drug Discovery, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

© The Author(s) 2021. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
A new module, membrANM, was developed for analysing membrane proteins, where the membrane is represented by a disk-shaped elastic network (Figure 1, lower left) (Leon and Bahar, 2012), and the force exerted by the membrane network is incorporated into the Hessian of the protein through a system-environment framework (see Supplementary Text).

2. **SignDy: Signature dynamics of protein families**

The SignDy module enables comparative analysis of the equilibrium dynamics of structural homologs and evaluation of their signature dynamics that often reflect their shared functional mechanisms (Zhang et al., 2019). The method is applicable to structural homologs that may share little sequence identity and/or exhibit functional diversity, as illustrated in Supplementary Figure 2 for 116 CATH superfamilies and the family of the periplasmic binding protein 1 (PBP-1) domains. The module evaluates the generic features shared by family members as well as specific features of subfamilies. This is made possible by (i) interfaces to various structural classification databases and servers for finding structure homologues (family members) given one input structure or ID; (ii) improved protein structure alignment protocols, including CEAlign (Shindyalov and Bourne, 1998) and automated chain matching procedures; (iii) optimal matching of normal modes accessible to family members; and (iv) comparative analyses using metrics such as covariance or mode-mode overlap. Residue fluctuation profiles and cross-correlations averaged over family members (Figure 1, lower right) define the signature dynamics, and deviations from the means describe their differentiation among family members. SignDy permits generation of dendrograms to cluster family members by their dynamics.

3. **CryoDy: Dynamics of Cryo-EM resolved structures**

CryoDy (Zhang et al., 2020) is designed to characterize the structural dynamics of cryo-EM resolved structures. It uses the topology-representing-network (TRN) algorithm to map electron densities associated with multiple residues to pseudo-atoms (ENM nodes), thus enabling efficient ENM-NMA and the use of low-resolution maps. The pipeline provides information on structural and dynamic properties including allosteric signal propagation paths based on existing ProDy tools and sampling of conformational landscapes through a new implementation of the adaptive ANM method, which works for both pseudoatomic and atomic models (see Figure 1, upper right). Its integration in ProDy permits a wealth of ENM-based analyses, in contrast to the powerful but more specialized tools in Scipion (de la Rosa-Trevin et al., 2016).

4. **ESSA: Essential Site Scanning Analysis**

ESSA (Kaynak et al., 2020) identifies essential residues. Essential residues are defined as those whose perturbation makes the highest impact (usually a shift to higher frequency) on the global modes intrinsically accessible to the system, being involved in biological activities (active or allosteric sites) or mechanical responses (hinges) (Supplementary Figure 3a-c). ESSA identifies these residues by evaluating the effect of increased crowding near each residue on the frequency dispersion of ENM modes. The change in global mode dispersion is measured by z-scores, which represent the mean shift in the frequency of the softest modes after pairwise matching between the original and perturbed models. ESSA integrates information on pocket geometry and local hydrophobic density data from Fpocket (Le Guilloux et al., 2009; Song et al., 2017) to provide an automated protocol for detecting allosteric pockets (Figure 1, upper left) and Supplementary Figure 3d.

Conclusion

Over the years, ProDy has been closing the gap in protein dynamics evaluations between theory and experiments. By virtue of its modular, object-oriented design and integration with scientific computing libraries, ProDy lends itself to easy development, scalability, and reproducibility. The features presented here extend its capabilities to analyse supramolecular systems resolved at low resolution (CryoDy), assess the conservation and differentiation of structural dynamics (SignDy), and identify essential sites that may impact the functional dynamics upon ligand binding (ESSA).

Acknowledgements. IMK thanks the Molecular Sciences Software Institute (MoSSI) and particularly Dr Andrew Abi-Mansour for guidance during fellowship.

Funding: This work has been supported by the National Institutes of Health (P41GM103712 to IB) and the MoSSI (COVID-19 Seed Software Fellowship to JK).

Conflict of Interest: none declared.

REFERENCES

Adams, P.D. et al. (2019) Announcing mandatory submission of PDB/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). Acta Crystallogr D Struct Biol, 75, 451-454.

Artigian, A.R. et al. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80, 505-515.

Bahar, I., Artigian, A.R. and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design, 2, 173-181.

Bahar, I., Jernigan, R.L. and Dill, K.A. (2017) Protein Actions: Principles and Modeling. Garland Science.

Bakan, A. et al. (2014) Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics, 30, 2681-2683.

Bakan, A., Meireles, L.M. and Bahar, I. (2011) ProDy: protein dynamics inferred from theory and experiments. Bioinformatics, 27, 1575-1577.

Cheng, A. et al. (2015) MRC2014: Extensions to the MRC format header for electron cryo-microscopy and tomography. J Struct Biol, 192, 146-150.

de la Rosa-Trevin, J.M. et al. (2016) Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. J Struct Biol, 195, 93-99.

Harris, C.R. et al. (2020) Array programming with NumPy. Nature, 585, 357-362.

Hunter, J.D. (2007) Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering, 9, 90-95.

Kaynak, B.T., Bahar, I. and Doruker, P. (2020) Essential Site Scanning Analysis: a new approach for detecting sites that modulate the dispersion of protein global motions. Computational and Structural Biotechnology Journal.

Lawson, C.L. et al. (2016) EMDDataBank unified data resource for 3DEM. Nucleic Acids Res, 44, D396-403.

Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics, 10, 168.

Leon, T.R. and Bahar, I. (2012) Constraints imposed by the membrane selectively guide the alternating access dynamics of the glutamate transporter GLUT1. Biophys J, 102, 1331-1340.

Li, H. et al. (2020) QuartaWeb: Integrated Chemical-Protein-Pathway Mapping for Polypharmacology and Chemogenomics. Bioinformatics, 36, 3935-3937.

Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incorporating combinatorial extension (CE) of the optimal path. Protein Eng, 11, 739-747.

Song, K. et al. (2017) Improved Method for the Identification and Validation of Allosteric Sites. J Chem Inf Model, 57, 2385-2363.

Virtanen, P. et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods, 17, 261-272.

Zhang, S. et al. (2019) Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity. Mol Biol Evol, 36, 2053-2068.
Zhang, Y. et al. (2020) State-dependent sequential allostery exhibited by chaperonin TRiC/CCT revealed by network analysis of Cryo-EM maps. *Prog Biophys Mol Biol*.  10.1016/j.pbiomolbio.2019.08.014