**ProDy: Protein Dynamics Inferred from Theory and Experiments**

Ahmet Bakan*, Lidio M. Meireles and Ivet Bahar*

Department of Computational and Systems Biology, and Clinical & Translational Science Institute, School of Medicine, University of Pittsburgh, 3064 BST3, 3501 Fifth Ave, Pittsburgh, PA 15213

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Summary: We developed a Python package, ProDy, for structure-based analysis of protein dynamics. ProDy allows for quantitative characterization of structural variations in heterogeneous datasets of structures experimentally resolved for a given biomolecular system, and for comparison of these variations with the theoretically predicted equilibrium dynamics. Datasets include structural ensembles for a given family or subfamily of proteins, their mutants and sequence homologues, in the presence/absence of their substrates, ligands or inhibitors. Numerous helper functions enable comparative analysis of experimental and theoretical data, and visualization of the principal changes in conformations that are accessible in different functional states. ProDy API has been designed so that users can easily extend the software and implement new methods.

Availability: ProDy is open-source and freely available under GNU General Public License from http://www.csb.pitt.edu/ProDy/.

1 INTRODUCTION

Protein dynamics plays a key role in a wide range of molecular events in the cell, including substrate/ligand recognition, binding, allosteric signaling, and transport. For a number of well-studied proteins, the Protein Data Bank (PDB) hosts multiple high-resolution structures. Typical examples are drug targets resolved in the presence of different inhibitors. These ensembles of structures convey information on the structural changes that are physically accessible to the protein, and the delineation of these structural variations provides insights into structural mechanisms of biological activity (Yang et al., 2008; Bakan & Bahar, 2009).

Existing computational tools and servers for characterizing protein dynamics are suitable for single structures (e.g., Anisotropic Network Model (ANM) server by Eyal et al., 2006), elNémo by Suhe & Sanejouand (2004), FlexServ by Camps et al., 2009), pairs of structures (e.g., open and closed forms of enzymes; MolMovDB by Gerstein & Krebs, 1998), or NMR models (e.g., PCA_NEST by Velazquez-Muriel et al., 2009; Raimondi et al., 2010). The lack of software for performing such operations is primarily due to the non-uniform content of structural datasets such as sequence variations at particular regions, including missing or substituted residues, short segments or loops. We developed ProDy to analyze and retrieve biologically significant information from such heterogeneous structural datasets. ProDy delivers information on the structural variability of target systems and allows for systematic comparison with the intrinsic dynamics predicted by theoretical models and methods, thus helping gain insight into the relation between structure, dynamics and function.

2 DESCRIPTION AND FUNCTIONALITY

2.1 Input for ProDy

The input for ProDy is the set of atomic coordinates in PDB format for the protein of interest, or simply the PDB id or sequence of the protein. Given a query protein, fast and flexible ProDy parsers are used to Blast search the PDB, retrieve the corresponding files (e.g., mutants, complexes, or sequence homologues with user-defined minimal sequence identity) from the PDB FTP server and extract their coordinates and other relevant data. Additionally, the program can be used to analyze a series of conformers from molecular dynamics (MD) trajectories inputted in PDB file format or programmatically through Python NumPy arrays. More information on the input format is given at the ProDy website tutorial and examples.

2.2 Protein ‘dynamics’ from experiments

The experimental data refer to ensembles of structures, X-ray crystallographic or NMR. These are usually heterogeneous datasets, in the sense that they have disparate coordinate data arising from sequence dissimilarities, insertions/deletions, or missing data due to unresolved disordered regions. In ProDy, we implemented algorithms for optimal alignment of such heterogeneous datasets and building corresponding covariance matrices. Covariance matrices describe the mean-square deviations in atomic coordinates from their mean position (diagonal elements) or the correlations between their pairwise fluctuations (off-diagonal elements). The principal modes of structural variation are determined upon principal component analysis (PCA) of the covariance matrix, as described previously (Bakan & Bahar, 2009).

*To whom correspondence should be addressed.*
2.3 Protein dynamics from theory and simulations

We have implemented classes for Gaussian network model (GNM) analysis and for normal mode analysis (NMA) of a given structure using the anisotropic network model (ANM) (Eyal et al., 2006). Both models have been widely used in recent years for analyzing and visualizing biomolecular systems dynamics (Bahar et al., 2010). The implementation is generic and flexible. The user can (i) build the models for any set of atoms, e.g., the substrate or inhibitor can be explicitly included to study the perturbing effect of binding on dynamics, and (ii) utilize user-defined or built-in distance-dependent or residue-specific force constants (Hinsen et al., 2000; Kovacs et al., 2004). ProDy also offers the option to perform essential dynamics analysis (EDA) (Amadei et al., 1993) of MD snapshots, which is equivalent to the singular value decomposition of trajectories to extract principal variations (Velazquez-Muriel et al., 2009).

2.4 Dynamics analysis example

Figure 1 illustrates the outputs generated by ProDy in a comparative analysis of experimental and computational data for p38 kinase (Bakan & Bahar, 2011). Panel A displays the dataset of 150 X-ray crystallographically resolved p38 structures retrieved from the PDB and optimally overlaid by ProDy. The ensemble contains the apo and inhibitor-bound forms of p38, thus providing information on the conformational space sampled by p38 upon inhibitor binding. Parsing structures, building and diagonalizing the covariance matrix to determine the principal modes of structural variations takes only 38 seconds on Intel CPU at 3.20 GHz. Panel C illustrate the first principal mode of structural variation (PC1; violet arrows) based exclusively on experimental structural dataset for p38.

As to generating computational data, two approaches are taken in ProDy: NMA of a representative structure using its ANM representation (panel B; color-coded such that red/blue regions refer to largest/smallest conformational mobilities); and EDA of MD trajectories provided that an ensemble of snapshots is provided by the user. The green arrows in panel C describe the first (lowest frequency, most collective) mode predicted by the ANM, shortly designated as ANM1. The heatmap in panel D shows the overlap (Marques & Sanejouand, 1995) between top-ranking (Marques & Sanejouand, 1995) between top-ranking PCA and ANM modes. The cumulative overlap between the top three pairs of modes is 0.73.

An important aspect of ProDy is the sampling of a representative set of conformers consistent with experiments - a feature expected to find wide utility in flexible docking and structure refinement. Panel E displays the conformational space sampled by experimental structures (blue dots), projected onto the subspace spanned by the top three PCA directions, which accounts for 59% of the experimentally observed structural variance. The conformations generated using the softest modes ANM1-ANM3 predicted to be intrinsically accessible to p38 in the apo form, are shown by the red dots. The sizes of the motions along these modes obey a Gaussian distribution with variance scaling with the inverse square-root of the corresponding eigenvalues. ANM conformers cover a subspace (green ellipsoidal envelope) that encloses all experimental structures. Detailed information on how to generate such plots and figures using ProDy is given in the online documentation, along with several examples of downloadable scripts.

2.5 Graphical Interface

We have designed a graphical interface, NMWiz, to enable users to perform ANM and PCA calculations from within a molecular visualization program. NMWiz is designed as a VMD (Humphrey et al., 1996) plugin, and is distributed within the ProDy installation package. It is used to do calculations for molecules loaded into VMD; and results are visualized on the fly. The plug-in allows for depicting color-coded network models and normal mode directions (Fig. 1B-C), displaying animations of various PCA and ANM modes, generating trajectories, and plotting square fluctuations.

2.6 Supporting features

ProDy comes with a growing library of functions to facilitate comparative analysis. Examples are functions to calculate, print and plot the overlaps between experiment, theory, and computations (Fig 1D) or to view the spatial dispersion of conformers (Fig 1E). For rapid and flexible analysis of large numbers of PDB structures, we designed a fast PDB parser. The parser can handle alternate locations and multiple models, and read specified chains or atom subsets selected by the user. We evaluated the performance of ProDy relative to Biopython PDB module (Hamelryck & Manderick, 2003) using 4701 PDB structures listed in the PDB SELECT dataset (Hobohm & Sander, 1994); we timed parsers for reading the PDB files and returning Cα-coordinates to the user (see documentation). The Python standard Biopython PDB parser eva-
Several web servers have been developed for characterizing protein dynamics, including eNémo (Suhre & Sanejouand, 2004), ANM (Eyal et al., 2006), and FlexServ (Camps et al., 2009). These servers perform coarse-grained ENM based NMA calculations, and as such are useful for elucidating structure-encoded dynamics of proteins. FlexServ also offers the option to use distance-dependent force constants (Kovacs et al., 2004), in addition to protocols for coarse-grained generation and PCA of trajectories. ProDy differs from these as it allows for systematic retrieval and comparative analysis of ensembles of heterogeneous structural datasets. Such datasets include structural data collected for members of a protein family in complex with different substrates/inhibitors. ProDy further allows for the quantitative comparison of the results from experimental datasets with theoretically predicted conformational dynamics. In addition, ProDy offers the following advantages: (i) it is extensible, interoperable, and suitable for use as a toolkit for developing new software, (ii) it provides scripts for automated tasks and batch analyses of large datasets, (iii) it has a flexible API suitable for testing new methods and hypotheses, and benchmarking them against existing methods with minimal effort and without the need to modify the source code, (iv) it allows for producing publication quality figures when used with Python plotting library Matplotlib; and (v) it provides the option to input user-defined distance-dependent force function or utilize elaborate classes that return force constants based on the type and properties of interacting residues (e.g., based on their secondary structure or sequence separation (Lezon & Bahar, 2010)).

3 DISCUSSION

Several web servers have been developed for characterizing protein dynamics, including eNémo (Suhre & Sanejouand, 2004), ANM (Eyal et al., 2006), and FlexServ (Camps et al., 2009). These servers perform coarse-grained ENM based NMA calculations, and as such are useful for elucidating structure-encoded dynamics of proteins. FlexServ also offers the option to use distance-dependent force constants (Kovacs et al., 2004), in addition to protocols for coarse-grained generation and PCA of trajectories. ProDy differs from these as it allows for systematic retrieval and comparative analysis of ensembles of heterogeneous structural datasets. Such datasets include structural data collected for members of a protein family in complex with different substrates/inhibitors. ProDy further allows for the quantitative comparison of the results from experimental datasets with theoretically predicted conformational dynamics. In addition, ProDy offers the following advantages: (i) it is extensible, interoperable, and suitable for use as a toolkit for developing new software, (ii) it provides scripts for automated tasks and batch analyses of large datasets, (iii) it has a flexible API suitable for testing new methods and hypotheses, and benchmarking them against existing methods with minimal effort and without the need to modify the source code, (iv) it allows for producing publication quality figures when used with Python plotting library Matplotlib; and (v) it provides the option to input user-defined distance-dependent force function or utilize elaborate classes that return force constants based on the type and properties of interacting residues (e.g., based on their secondary structure or sequence separation (Lezon & Bahar, 2010)).

4 CONCLUSION

ProDy is a free, versatile, easy-to-use and powerful tool for inferring protein dynamics from both experiments (i.e. PCA of ensembles of structures) and theory (i.e. GNM, ANM and EDA of MD snapshots). ProDy complements existing tools by allowing the systematic retrieval and analysis of heterogeneous experimental datasets, leveraging on the wealth of structural data deposited in the PDB to gain insights into structure-encoded dynamics. In addition, ProDy allows for comparison of the results from experimental datasets with theoretically predicted conformational dynamics. Finally, through a flexible Python-based API, ProDy can be used to quickly test and implement new methods and ideas, thus lowering the technical barriers to apply such methods in more complex computational analyses.

ACKNOWLEDGEMENTS

Funding: Support from NIH IR01GM086238-01 is acknowledged by IB.

REFERENCES