BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 18 2014, pages 2681–2683
doi:10.1093/bioinformatics/btu336

Structural bioinformatics

Advance Access publication May 21, 2014

Evol and *ProDy* for bridging protein sequence evolution and structural dynamics

Ahmet Bakan†, Anindita Dutta†, Wenzhi Mao, Ying Liu, Chakra Chennubhotla, Timothy R. Lezon and Ivet Bahar*

Department of Computational and Systems Biology, and Clinical & Translational Science Institute, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA

Associate Editor: Anna Tramontano

ABSTRACT

Correlations between sequence evolution and structural dynamics are of utmost importance in understanding the molecular mechanisms of function and their evolution. We have integrated *Evol*, a new package for fast and efficient comparative analysis of evolutionary patterns and conformational dynamics, into *ProDy*, a computational toolbox designed for inferring *pro*tein *dyn*amics from experimental and theoretical data. Using information-theoretic approaches, *Evol* coanalyzes conservation and coevolution profiles extracted from multiple sequence alignments of protein families with their inferred dynamics.

Availability and implementation: *ProDy* and *Evol* are open-source and freely available under MIT License *from* http://prody.csb.pitt.edu/.

Contact: bahar@pitt.edu

Received on December 24, 2013; revised on April 20, 2014; accepted on May 8, 2014

1 INTRODUCTION

The significance of protein dynamics in a wide range of biological functions, including cell signaling, regulation and machinery is widely established (Bahar *et al*., 2010; Bhabha *et al*., 2011; Marsh *et al*., 2012). In many cases, sequence variability goes hand in hand with structural dynamics (Glembo *et al*., 2012; Liu and Bahar, 2012; Marks *et al*., 2011; Micheletti, 2012; Worth *et al*., 2009; Zheng *et al*., 2005). Structural dynamics correlates with evolvability (Tokuriki and Tawfik, 2009) or sequence and conformational diversity (Friedland *et al*., 2009) and enables adaptation to substrate binding while maintaining specificity (Liu *et al*., 2010). To our knowledge, existing software usually relate evolutionary properties to *static* structures (Ashkenazy *et al*., 2010; Morgan *et al*., 2006; Wainreb *et al*., 2011), or they are exclusively dedicated to either sequence analysis (Waterhouse *et al*., 2009) or structural dynamics (Eyal *et al*., 2006; Suhre and Sanejouand, 2004). There is a need for methods that allow combined analysis of sequence (co)evolution and structural dynamics. These would be particularly useful if they could be performed and visualized in a versatile, integrated computing environment.

Toward addressing this need, we introduce the v1.5 of *ProDy* (Bakan *et al*., 2011) with *Evol* applications. Highlights of the new version are rich methods for coevolutionary analysis, and extensions for analyzing and interpreting structural dynamics, following the approach adopted in our recent comparative study of sequence conservation and coevolution patterns versus structure/dynamics properties for a representative set of protein families (Liu and Bahar, 2012), which has been validated in detailed case studies (e.g. General *et al*., 2014; Liu *et al*., 2010). A distinctive feature of *ProDy* is its capability to extract mechanistic information from principal component analysis (PCA) of ensembles of structures (e.g. drug targets) (Bakan and Bahar, 2009). The new release has several new modules and command line applications named '**evol**' to evaluate sequence conservation and coevolution using information-theoretic and statistical approaches. To our knowledge, this is the only package that enables comparative analysis of protein dynamics with sequence evolution data extracted from multiple sequence alignments (MSAs) for protein families.

2 DESCRIPTION AND FUNCTIONALITY

2.1 Input for *ProDy* and *Evol*

The input for *ProDy* is a set of protein coordinates in PDB format, or simply the PDB ID or protein sequence. The speed of PDB parser and AtomGroup classes has been increased in the current version, such that parsing coordinates is 4.5–40 times faster than using Biopython PDB module (Hamelryck and Manderick, 2003), and atomic data storage occupies 10 times less memory footprint. We implemented efficient and flexible features for handling MSAs. Notably, the new MSA parser can evaluate various formats at a rate of 700 MB/s (on 3.6 GHz Intel Xeon CPU, 16 GB RAM and Samsung SSD) and is up to 80 times faster than the alignment parser of Biopython (Cock *et al*., 2009). Flexible classes store MSA data parsimoniously in the memory and provide ways of subsampling. Sequences can be filtered based on their labels to retain those in certain categories (e.g. human) and sliced to retain specific regions or sequences (e.g. regions matching structurally resolved amino acids). Such refinements, performed in a fraction of a second, allow for real-time processing of large MSAs and systematic analyses of protein families.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

2.2 Coevolution analysis

Evol offers a rich set of features for evaluating and plotting evolutionary properties of amino acids. Methods based on mutual information (Dunn *et al.*, 2008), statistical coupling analysis (SCA) recent extension (Halabi *et al.*, 2009), observed-minus-expected-squared covariance (Kass and Horovitz, 2002) and direct information (DI) (Marks *et al.*, 2011; Weigt *et al.*, 2009) have been implemented for coevolution analysis. Our implementations of these methods follow the descriptions in their respective papers. We rigorously tested our methods, cross-checking our results with the code that came with the cited papers. *Evol* can operate in turbo mode when there is sufficient memory (twice the size of MSA file); otherwise it falls back to a memory efficient mode. Benchmarking the performance of different implementations also show that, *Evol* algorithms written in C/Python run 1.5 (DI) to 7 (SCA) times faster than the original implementations in Matlab. Furthermore, *Evol* takes account of ambiguous (e.g. Asx) and modified (e.g. selenocysteine, pyrrolysine) amino acids or gaps. More specific requirements, such as the occupancy of amino acid positions, can also be satisfied using preprocessing methods described in the previous section. Minimal numbers of sequences to be included in the MSAs are recommended to be 100 and 250 in SCA and DI methods, respectively, in accord with the original studies. All methods are available in the API and through '**evol**' program, and their usage is illustrated in the *Evol* Tutorial on the *ProDy* Web site.

2.3 Structure and dynamics analysis

ProDy was originally designed for inferring structural dynamics from PCA of experimental structural datasets, as well as predictions of the Gaussian network model (GNM) of other elastic network models (ENMs) (Bahar *et al.*, 2010). Building on these methods, we have implemented perturbation-response scanning (Atilgan and Atilgan, 2009), an ENM variant with structure-based force constants (Lezon and Bahar, 2010), rotations-translations of blocks method (Tama *et al.*, 2000), membrane ENM model (Lezon and Bahar, 2012) and ENM reduction (Hinsen *et al.*, 2000) and extension algorithms that enable mapping the model to smaller or larger parts of the studied system. In addition, we added features for essential dynamics analysis (EDA) (Amadei *et al.*, 1993) of MD trajectories. Along with the *Evol* suite, *ProDy* now permits comparison of sequence evolution data and structural dynamic patterns predicted by ENMs or deduced from experimental data (PCA) or simulations (EDA).

2.4 Comparisons of sequence evolution and structural dynamics

Of particular interest is to understand the dynamical properties of conserved amino acids and *vice versa*. On calculation of mobility and conservation profiles for a given protein or a protein family, the profiles can be compared using Pearson's or Spearman's correlation coefficients (Liu and Bahar, 2012). *ProDy* and *Evol* API functions enable such comparisons by facilitating mapping between structure- and sequence-based models, i.e. missing residues in the structure or sequence are represented with dummy atoms, and outputting results as numerical arrays that can be fed directly into the statistical analysis modules of SciPy, NumPy and Matplotlib.

2.5 *NMWiz* for visual comparative analysis

We enhanced in v1.5 the capabilities of the *Normal Mode Wizard* (*NMWiz*) plug-in, which is now distributed with VMD (Humphrey *et al.*, 1996). *NMWiz* can be used to analyze all molecule and trajectory file formats supported by VMD and to perform a comparative visual analysis of structural dynamics and sequence evolution. Figure 1 displays screenshots of VMD molecular representations (Panel B) and MultiPlot and Heatmapper plots (Panels C and D) showing conservation and mobility profiles and evolutionary and dynamical correlations, all generated through *NMWiz*.

2.6 An illustrative example

Figure 2 illustrates an application of *ProDy* and *Evol* to compare the sequence conservation and coevolution patterns of the RNase A family of proteins with the global dynamics of a structurally resolved (Holloway *et al.*, 2009) member of the family. Panel A shows the correlation between sequence entropy (*gray bars*) and mobility profile of residues predicted by the GNM based on all modes (*black*), and a subset of global modes (eight lowest frequency modes, *blue*). Active site residues Q11, K41 and H119 have minimal entropy and low mobility. Panel B displays the ribbon diagrams color-coded by residue conservation (*left*) and intrinsic conformational mobility (*right*). Highly conserved (low entropy) residues, colored *blue* on the *left* diagram also have lower mobility (*blue*, *right*). Conversely, highly variable residues (*red*, *left*) tend to occupy highly mobile regions (*red*, *right*). A few residues are highlighted (encircled in A and B) to ease the comparison. Panel C shows the mutual information map generated for the family. The bright points (*cyan* to *red*) in the heat map refer to pairs that have high coevolution propensities. A number of evolutionarily correlated but sequentially distant (≥6 intervening residues) pairs of sites are highlighted (*circles*), including spatially close (*magenta*) or distant (*orange*) pairs shown in panel D. Notably, (C65, C72) forms a disulfide bridge; (T82, H48) make side chain (polar) interactions (*left* diagram); and the pairs (N71, Q11) and (T36, D14) are

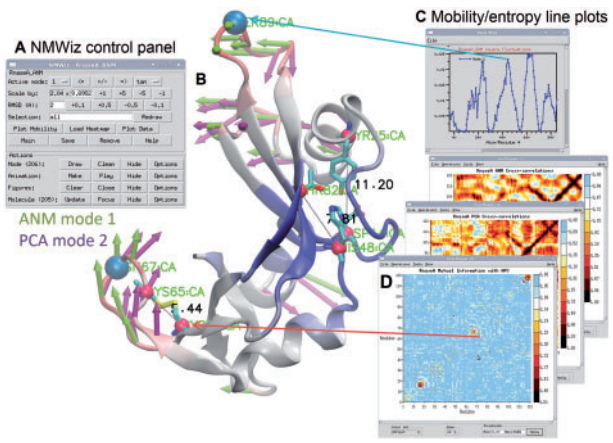

Fig. 1. *NMWiz* for comparative analysis of *ProDy* and *Evol* output. (**A**) *NMWiz* control panel. (**B**) Protein and normal mode representations, (**C**) mobility and conservation profiles and (**D**) cross-correlations in dynamics and coevolution generated via *NMWiz*

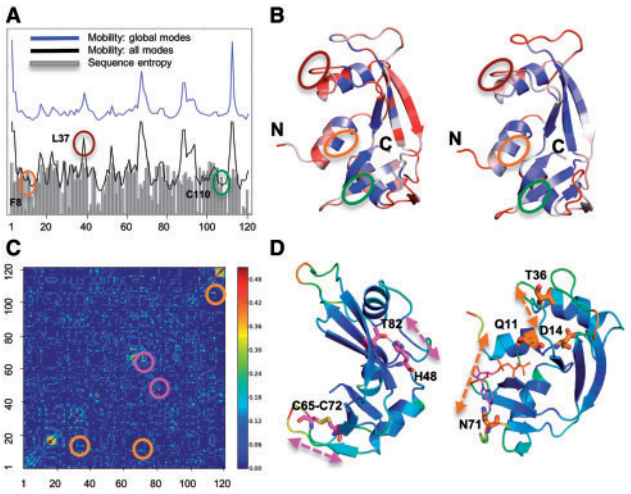

Fig. 2. Comparison of sequence conservation and residue mobility (panels **A** and **B**), and sequence-coevolution and spatial location of selected coevolving pairs (panels **C** and **D**) for RNase A. See text in Section 2.6 for more details

presumably involved in allosteric interactions (*right* diagram). The *right* diagram in panel D displays the RNase A crystallized in the presence of an inhibitor-like substrate (*thin stick* representation) (Holloway *et al.*, 2009). Q11 and N71 form hydrogen bonds with the substrate to ensure binding/recognition specificity, whereas D14 (near the binding site) shows long-range coevolution with a distant part of the residue (T36) suggestive of allosteric communication.

3 CONCLUSION

Evol adds new API features and command line applications to *ProDy* for rapid assessment and visualization of sequence conservation and coevolution patterns and allows for examining these results in the light of the structure and dynamics of proteins, motivated by our current understanding of the role of intrinsic dynamics in sequence evolution. The *ProDy* API and the new extensions implemented here can harness efficient and powerful features of other open-source scientific packages (e.g. NumPy, SciPy and Matplotlib), to harness their efficient and powerful features, thus making the API suitable for interactive usage and rapid and easy development of new applications.

Funding: The work was supported by National Institutes of Health [5R01GM099738 and P41 GM103712 to I.B.] and fellowship by Tsinghua University [to W.M.].

Conflict of Interest: none declared.

REFERENCES

Amadei,A. *et al.* (1993) Essential dynamics of proteins. *Proteins*, **17**, 412–425.
Ashkenazy,H. *et al.* (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
Atilgan,C. and Atilgan,A.R. (2009) Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.*, **5**, e1000544.
Bahar,I. *et al.* (2010) Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.*, **39**, 23–42.
Bakan,A. and Bahar,I. (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl Acad. Sci. USA*, **106**, 14349–14354.
Bakan,A. *et al.* (2011) ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics*, **27**, 1575–1577.
Bhabha,G. *et al.* (2011) A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, **332**, 234–238.
Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.
Eyal,E. *et al.* (2006) Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, **22**, 2619–2627.
Friedland,G.D. *et al.* (2009) A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput. Biol.*, **5**, e1000393.
General,I.J. *et al.* (2014) ATPase subdomain IA is a mediator of interdomain allostery in Hsp70 molecular chaperones. *PLoS Comput. Biol.*, **10**, e1003624.
Glembo,T.J. *et al.* (2012) Collective dynamics differentiates functional divergence in protein evolution. *PLoS Comput. Biol.*, **8**, e1002428.
Halabi,N. *et al.* (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**, 774–786.
Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
Hinsen,K. *et al.* (2000) Harmonicity in slow protein dynamics. *Chem. Phys.*, **261**, 25–37.
Holloway,D.E. *et al.* (2009) Influence of naturally-occurring 5′-pyrophosphate-linked substituents on the binding of adenylic inhibitors to ribonuclease A: an X-ray crystallographic study. *Biopolymers*, **91**, 995–1008.
Humphrey,W. *et al.* (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.
Lezon,T.R. and Bahar,I. (2010) Using entropy maximization to understand the determinants of structural dynamics beyond native contact topology. *PLoS Comput. Biol.*, **6**, e1000816.
Lezon,T.R. and Bahar,I. (2012) Constraints imposed by the membrane selectively guide the alternating access dynamics of the glutamate transporter Glt Ph. *Biophys. J.*, **102**, 1331–1340.
Liu,Y. and Bahar,I. (2012) Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.*, **29**, 2253–2263.
Liu,Y. *et al.* (2010) Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs. *PLoS Comput. Biol.*, **6**, 15.
Marks,D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.
Marsh,J.A. *et al.* (2012) Probing the diverse landscape of protein flexibility and binding. *Curr. Opin. Struct. Biol.*, **22**, 643–650.
Micheletti,C. (2012) Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. *Phys. Life Rev.*, **10**, 1–26.
Morgan,D.H. *et al.* (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, **22**, 2049–2050.
Suhre,K. and Sanejouand,Y.-H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, **32**, W610–W614.
Tama,F. *et al.* (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, **41**, 1–7.
Tokuriki,N. and Tawfik,D.S. (2009) Protein dynamism and evolvability. *Science*, **324**, 203–207.
Wainreb,G. *et al.* (2011) Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics*, **27**, 3286–3292.
Waterhouse,A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.
Worth,C.L. *et al.* (2009) Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.*, **10**, 709–720.
Zheng,W. *et al.* (2005) Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. *Structure*, **13**, 565–577.