Bioinformatics, 35(22), 2019, 4824–4826
doi: 10.1093/bioinformatics/btz460
Advance Access Publication Date: 4 June 2019
Applications Note

Structural bioinformatics

WDSPdb: an updated resource for WD40 proteins

Jing Ma[1,†], Ke An[1,†], Jing-Bo Zhou[1], Nuo-Si Wu[2], Yang Wang[3,4], Zhi-Qiang Ye[1,*] and Yun-Dong Wu[1,5,*]

[1]Lab of Computational Chemistry and Drug Design, State Key Laboratory of Chemical Oncogenomics, Peking University Shenzhen Graduate School, Shenzhen 518055, China, [2]College of Information Engineering, Shenzhen University, Shenzhen 518060, China, [3]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [4]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA and [5]College of Chemistry, Peking University, Beijing 100871, China

*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.

Associate Editor: Alfonso Valencia

Received on December 24, 2018; revised on April 28, 2019; editorial decision on May 27, 2019; accepted on May 29, 2019

Abstract

Summary: The WD40-repeat proteins are a large family of scaffold molecules that assemble complexes in various cellular processes. Obtaining their structures is the key to understanding their interaction details. We present WDSPdb 2.0, a significantly updated resource providing accurately predicted secondary and tertiary structures and featured sites annotations. Based on an optimized pipeline, WDSPdb 2.0 contains about 600 thousand entries, an increase of 10-fold, and integrates more than 37 000 variants from sources of ClinVar, Cosmic, 1000 Genomes, ExAC, IntOGen, cBioPortal and IntAct. In addition, the web site is largely improved for visualization, exploring and data downloading.

Availability and implementation: http://www.wdspdb.com/wdsp/ or http://wu.scbb.pkusz.edu.cn/wdsp/.

Contact: yezq@pkusz.edu.cn or wuyd@pkusz.edu.cn

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

1 Introduction

The WD40-repeat proteins are a subfamily of β-propellers, and their sequence and structure relationships and association with diseases have been widely studied (Kopec and Lupas, 2013; Paoli, 2001; Pons *et al.*, 2003; Song *et al.*, 2017). As one of the most popular interactors in protein–protein interaction (PPI) networks, they act as scaffolds to assemble various molecular machineries, and play versatile roles in fundamental biological processes including signal transduction, ubiquitination, cell cycle control, etc. (Stirnimann *et al.*, 2010; Xu and Min, 2011). Obtaining their structural information is the key to revealing their interacting details and thus to understanding their biological functions and to obtaining insights to the underlying pathogenic mechanisms, but available experimental structures are heavily lacked regarding their abundance in eukaryotic proteomes.

WDSPdb (Wang *et al.*, 2015) is a database providing accurate structure predictions and featured sites annotations specifically for WD40 domains, based on the WDSP tool (Wang *et al.*, 2013). WD40 domains, as a type of β-propellers, are composed of several repeated β-sheet units with a circular layout. WDSPdb offers the boundaries of β-strands for each repeat unit, and affords thermal-stabilizing hydrogen bond network sites and potential interaction hotspots. These data are deficient in general-purpose domain databases, but are indispensable to understand the functional roles of WD40 proteins. Since its publication, WDSPdb 1.0 has served the scientific community frequently. However, its contents are currently heavily lagged compared to the rapid increase of protein sequences in public databases, and the data coverage is relatively small due to its over-strict criteria of data inclusion. In this work, we have optimized the overall curation pipeline, and then applied it to a more recent version of UniProtKB (The UniProt Consortium, 2017) to construct WDSPdb 2.0.

2 Materials and methods

We first updated the WDSP tool, which relies on a WD40-specific position weight matrix (PWM) and PSIPRED (Buchan *et al.*, 2013) as backends. We have expanded the experimental structures from 33 to 65 for generating the PWM. Meanwhile, the PSIPRED has been replaced from V3 to V4. With these improvements, the updated WDSP has outperformed its previous version and other general-purpose domain annotations, as measured by the 5-fold cross-validation F1 score (Supplementary Material).

We have then optimized the overall annotation pipeline for more comprehensive inclusion of WD40 proteins and better annotation quality (Fig. 1A and Supplementary Material). The optimized pipeline is briefly described as follows: (i) retrieve the protein sequences of UniProtKB, including Swiss-Prot and TrEMBL section. (ii) Utilize the HMMSearch to screen all the input sequences based on 54 WD40-related profiles, and retain the sequences with E-value no greater than 10 as WD40 candidates. (iii) Employ the updated WDSP to predict the secondary structures and the featured sites. (iv) Assign confidence categories (‘High’, ‘Middle’ and ‘Low’) to each WD40 candidate according to our customized rules. (v) Use MODELLER (Webb and Sali, 2016) to build 3D structures for single-domain WD40s, and integrate missense variants and their associated annotations to WD40s with ‘High’ confidence from the Swiss-Prot section.

WDSPdb 2.0 is based on UniProtKB (release July 5, 2017), a more recent version, and an optimized curation pipeline allowing the inclusion of more non-canonical WD40 proteins. As a result, the data coverage is about 10 times of WDSPdb 1.0. In brief, it contains 594 319 WD40 proteins with 4 033 034 repeats from 4426 species. Among these proteins, 852 295 potential side-chain hydrogen bond networks and 4 963 216 PPI hotspots were predicted. Specifically, from ClinVar (Landrum *et al.*, 2018), Cosmic (Forbes *et al.*, 2017), IntOGen (Gonzalez-Perez *et al.*, 2013), cBioPortal (Cerami *et al.*, 2012), IntAct (Kerrien *et al.*, 2012), 1000 Genomes (1000 Genomes Project Consortium, 2015) and ExAC (Exome Aggregation Consortium, 2016), we have mapped to 252 WD40 proteins 37 184 variants, which are pathogenic, cancer-related, cancer-driver, cancer highly recurrent, PPI-influencing or neutral. WDSPdb 2.0 comprises almost all of the entries in WDSPdb 1.0, and only a few entries are exclusive in WDSPdb 1.0 due to entry merging, removing and renaming in the process of UniProtKB updates (Fig. 1B and Supplementary Material). As expected, the intersection of WDSPdb 2.0 and 1.0 mainly belongs to the ‘High’ confidence category, and most newly added entries are assigned to other confidence categories, since the new pipeline has adopted looser inclusion criteria. Many proteins that are widely considered as WD40 proteins but absent in WDSPdb 1.0 have been included in WDSPdb 2.0, such as LRRK2, PALB2 and APAF1. Taken together, WDSPdb 2.0 is much more comprehensive regarding the record number and annotation information.

We re-implemented the web interface using Django to provide cleaner and more organized browsing experiences. It adopts a powerful table plug-in that enables customized data display and download in multiple formats, and has replaced the visualization tool to NGL viewer (Rose and Hildebrand, 2015) for faster loading and smoother operation. A REST service has also been implemented for downloading the secondary structure annotations. In addition, we deployed the updated WDSP tool with options of parameter tuning (the searching database and the iterative times), which would provide predictions for users’ own sequences.

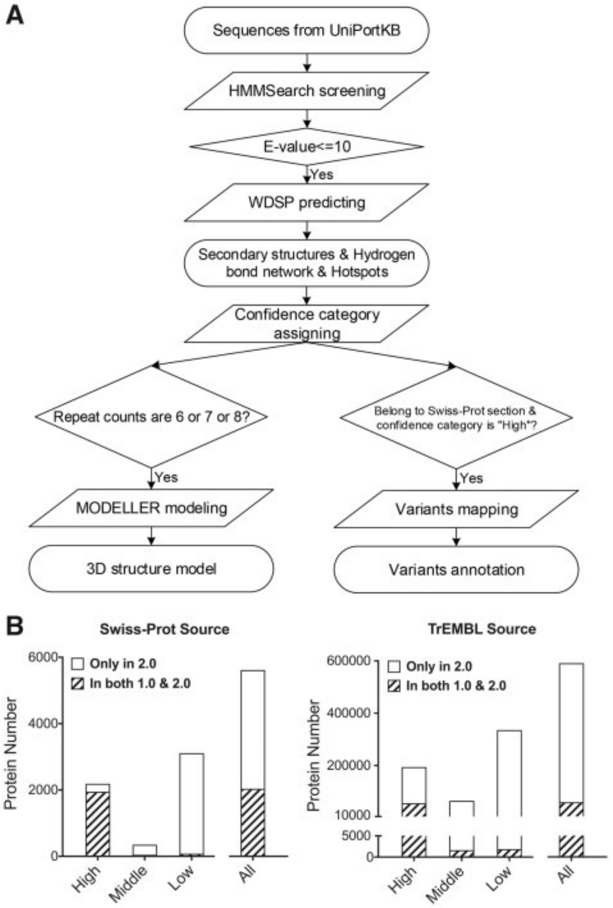

Fig. 1. (**A**) The pipeline of curating WDSPdb 2.0. (**B**) The distribution of WD40 proteins in different confidence categories, calculated separately for Swiss-Prot and TrEMBL source. The WD40 proteins in both WDSPdb 1.0 and 2.0 are indicated with slashes. ‘All’ equals the sum of ‘High’, ‘Middle’ and ‘Low’

3 Conclusion and discussion

WDSPdb 2.0 has incorporated significant improvements. The version 1.0 is confined to typical WD40 proteins only, but users have frequently requested annotations of atypical ones. This update recorded as many as possible putative WD40 proteins with more accurate structure predictions, and has assigned confidence levels to meet requirements of customized usages. The integration of variant data will enable the direct and intuitive exploring of the relationship between variants and featured sites in the structural context. The web interface is also largely enhanced for better browsing, visualization, and downloading. We will regularly update WDSPdb to continuously benefit the researchers in the fields of repeat proteins, PPIs and genetic variants interpretation.

Acknowledgements

The authors would like to thank all the WDSPdb users for their helpful feedback and suggestions. They also thank the anonymous reviewers for the valuable advice of improving the manuscript and the web server.

Funding

This work was supported by the National Natural Science Foundation of China [31471243]; and the Shenzhen Basic Research Program [JCYJ20170818085409785 and JCYJ20170412150507046].

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Buchan,D.W. *et al*. (2013) Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res*., **41**, W349–W357.

Cerami,E. *et al*. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*., **2**, 401–404.

Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

Forbes,S.A. *et al*. (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, **45**, D777–D783.

Gonzalez-Perez,A. *et al*. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.

Kerrien,S. *et al*. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res*., **40**, D841–846.

Kopec,K.O. and Lupas,A.N. (2013) Beta-propeller blades as ancestral peptides in protein evolution. *PLoS One*, **8**, e77074.

Landrum,M.J. *et al*. (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*., **46**, D1062–D1067.

Paoli,M. (2001) Protein folds propelled by diversity. *Prog. Biophys. Mol. Biol*., **76**, 103–130.

Pons,T. *et al*. (2003) Beta-propellers: associated functions and their role in human diseases. *Curr. Med. Chem*., **10**, 505–524.

Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res*., **43**, W576–W579.

Song,R. *et al*. (2017) Disease association and druggability of WD40 repeat proteins. *J. Proteome Res*., **16**, 3766–3773.

Stirnimann,C.U. *et al*. (2010) WD40 proteins propel cellular networks. *Trends Biochem. Sci*., **35**, 565–574.

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **45**, D158–D169.

Wang,Y. *et al*. (2015) WDSPdb: a database for WD40-repeat proteins. *Nucleic Acids Res*., **43**, D339–344.

Wang,Y. *et al*. (2013) A method for WD40 repeat detection and secondary structure prediction. *PLoS One*, **8**, e65705.

Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics*, **54**, 5.6.1–5.6.37.

Xu,C. and Min,J. (2011) Structure and function of WD40 domain proteins. *Protein Cell*, **2**, 202–214.