Identification of WD40 repeats by secondary structure-aided profile–profile alignment

Chuan Wang\textsuperscript{a,b}, Xiaobao Dong\textsuperscript{a}, Lei Han\textsuperscript{c}, Xiao-Dong Su\textsuperscript{d}, Ziding Zhang\textsuperscript{a,e}, Jinyan Li\textsuperscript{c,e}, Jiangning Song\textsuperscript{f,g,h,***}

\textsuperscript{a} State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China
\textsuperscript{b} Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA
\textsuperscript{c} Center for Cancer Molecular Diagnosis, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, China
\textsuperscript{d} State Key Laboratory of Protein and Plant Gene Research and Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, China
\textsuperscript{e} Advanced Analytics Institute and Centre for Health Technologies, University of Technology Sydney, 81 Broadway, Sydney, NSW 2007, Australia
\textsuperscript{f} National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
\textsuperscript{g} Infection and Immunity Program, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia
\textsuperscript{h} Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

HIGHLIGHTS

- We develop an accurate WD40 repeat prediction method based on secondary structure and a profile–profile alignment.
- A novel alignment scoring function that combines dot product and BLOSUM62 is designed.
- The WDRR web server and the datasets are available at http://protein.cau.edu.cn/wdrr/.

ARTICLE INFO

Article history:
Received 24 July 2015
Received in revised form
20 January 2016
Accepted 16 March 2016
Available online 25 March 2016

Keywords:
WD40 repeat
Profile–profile alignment
Remote sequence homology
Secondary structure prediction
Bioinformatics

ABSTRACT

A WD40 protein typically contains four or more repeats of ~40 residues ended with the Trp-Asp dipeptide, which folds into β-propellers with four β strands in each repeat. They often function as scaffolds for protein–protein interactions and are involved in numerous fundamental biological processes. Despite their important functional role, the “velcro” closure of WD40 propellers and the diversity of WD40 repeats make their identification a difficult task. Here we develop a new WD40 Repeat Recognition method (WDRR), which uses predicted secondary structure information to generate candidate repeat segments, and further employs a profile–profile alignment to identify the correct WD40 repeats from candidate segments. In particular, we design a novel alignment scoring function that combines dot product and BLOSUM62, thereby achieving a great balance of sensitivity and accuracy. Taking advantage of these strategies, WDRR could effectively reduce the false positive rate and accurately identify more remote homologous WD40 repeats with precise repeat boundaries. We further use WDRR to re-annotate the Pfam families in the β-propeller clan (CL0186) and identify a number of WD40 repeat proteins with high confidence across nine model organisms. The WDRR web server and the datasets are available at http://protein.cau.edu.cn/wdrr/.

1. Introduction

WD40 repeat proteins are well-known to function as scaffolds for protein–protein interactions and play diverse functional roles in cellular processes, such as regulatory pathways of cell cycle (Yu, 2007), cell apoptosis (Adams and Cory, 2002; Yuan et al., 2010; Reubold et al., 2011), autophagy (Fimia et al., 2007), gene transcription (Pickles et al., 2002), signal transduction (Datta and

* Corresponding authors.
** Corresponding author at: Monash Centre for Data Science, Faculty of Information Technology, Monash University, VIC 3800, Australia.
E-mail addresses: cwan@carnegiescience.edu (C. Wang), yndxb@126.com (X. Dong), hanlei_45@126.com (L. Han), xdsu@pku.edu.cn (X.-D. Su), zidingzhang@cau.edu.cn (Z. Zhang), jinyan.Li@uts.edu.au (J. Li), jiangning.song@monash.edu (J. Song).
Moses, 2000; Chen et al., 1995), histone modification (Song et al., 2008; Suganuma et al., 2008), DNA damage repair (Scrima et al., 2008), RNA modification (Ren et al., 2011; Vander Kozi et al., 2010; Ajith et al., 2001), vesicular traffic (Stagg et al., 2007), cytoskeleton assembly (Hartman et al., 1998; Chan et al., 2011), and chromatin assembly (Li and Luan, 2011). Mutations in WD40 repeats could cause diseases like amelogenesis imperfecta (El-Sayed et al., 2009), severe cerebral cortical malformations (Bilgiovâr et al., 2010), and primary open-angle glaucoma (Gallenberger et al., 2011). WD40 proteins themselves do not possess any known catalytic activity. Although a few are found in bacteria, WD40-containing proteins are primarily abundant in eukaryotes. About one percent of the proteins themselves do not have 7 or 14 WD40 repeats. Despite the distribution of WD40 repeats in proteins. Although this could maximize the ability to find all repeats in a protein, it can also miss proteins that do not have 7 or 14 WD40 repeats. Despite the availability of these tools, in the latest reviews and analysis of WD40 proteins (Stirnimann et al., 2010; Smith, 2008; Xu and Min, 2011; Mishra et al., 2014), the authors still utilized BLAST (Altschul et al., 1997), Pfam (Finn et al., 2010) and SMART (Letunic et al., 2009) to identify WD40 proteins.

Smith (2008) made an excellent discussion about the challenge of identifying WD40 repeats and proposed a basic procedure. In addition to the strand shift, the number of repeats varies within different WD40 proteins, and the repeats often form more than one β-propeller in a single protein chain. Moreover, WD40 proteins often have short motifs or other domains inserted within WD40 repeats or between propellers. In light of these difficulties, Smith (2008) suggested that an iterative approach should be used, which first identified repeats that were easy to find, and then examined the rest of the proteins for more divergent repeats.

In this paper, we used a different WD40 repeat identification procedure from a reverse perspective, which first identifies potential WD40 repeats as many as possible, then filters these repeats according to the criteria of the basic structural assumptions of WD40 domains (Smith, 2008). In our procedure, first, we construct a WD40 repeat consensus sequence and profile from available structures; second, the predicted secondary structure is used to partition the query sequence into candidate segments; third, we incorporate a profile–profile alignment algorithm with a combined scoring function of dot product and BLOSUM62 to align the candidate segments to the WD40 repeat sequence profile; finally, we combine the alignment score with a secondary structure similarity measure to sort and select tentative WD40 repeats.
from the candidate segments. This approach is termed as WD40 Repeat Recognition (WDRR). The flowchart of WDRR is shown in Fig. 2.

We tested the performance of WDRR on a benchmark dataset generated from the SCOP database (version 1.75) (Murzin et al., 1995) and compared its performance with those of the sequence – HMM search method HMMER3 (Eddy, 2009) used by Pfam 25.0 (Finn et al., 2010) and the HMM-HMM search method HHsearch (Söding, 2005). All results indicate that WDRR provides a better performance for identifying more reliable WD40 repeats with more accurate boundaries. Performance comparison between WDRR and WDSP suggests that WDRR is more sensitive than WDSP and is an alternatively useful WD40 repeat identification method. To explore WDRR’s real-world application, we then applied WDRR to re-annotate the Pfam WD40 family (PF00400), as well as all DUF (Domain of Unknown Function) members in the β-propeller clan (CL0186). Moreover, we further performed a proteome-wide analysis of WD40 repeat proteins across nine model organisms and identified a number of high-confidence WD40 repeat proteins in each organism.

2. Materials and methods

2.1. WD40 repeat profile construction

To construct the WD40 repeat sequence profile to be used as the template in alignment-based recognition, we collected all the available WD40 structures from SCOP 1.75 (Murzin et al., 1995) and the Protein Data Bank (PDB) (Berman et al., 2000). The WD40 repeat segments of these structures, which start from d-strand and end with the WD dipeptide (or the corresponding residues) at the C-terminal of c-strand, were extracted manually. To reduce the sequence redundancy of the WD repeats, only those segments that share less than 40% identity between each other were kept. Finally, a total of 86 WD40 repeat segments were obtained.

We then used Matt v1.00 (Menke et al., 2008) to generate a multiple structure alignment to these structure segments and refined the output alignment with ClustalW 2.0 (Larkin et al., 2007). The consensus sequence of the refined alignment was considered as the template of the WD40 repeat sequence. An amino acid frequency profile with the Henikoff and Henikoff weights (Henikoff and Henikoff, 1994) were calculated and used as the template of the WD40 repeat profile. The secondary structures of the segments were calculated using DSSP (Kabsch and Sander, 1983) and mapped to the refined alignment to generate the consensus secondary structure sequence of the profile.

2.2. Query sequence preparation using predicted secondary structure

For a query protein sequence, PSI-BLAST (Altschul et al., 1997) search with three iterations and an e-value cutoff of 0.001 against the NCBI nr90 database (Yan et al., 2009) was performed to generate the multiple sequence alignment (MSA). The frequency profile of the query sequence was calculated from the MSA using the same approach in calculating the template profile. Meanwhile, we employed PSIPRED v3.2 (Jones, 1999) to predict the secondary structure of the query sequence.

Since the template WD40 sequence is only composed of ~40 residues, the local alignment against full-length protein sequences based on the Smith–Waterman algorithm (Smith and Waterman, 1981) is usually used, which often misses terminal residues at boundaries. To avoid this issue, we used a four-β-strand window to slide along the query sequence according to the predicted secondary structure. Each four-β-strand segment was considered as a candidate to be further aligned against the template WD40 profile using global alignment. For example, if a query sequence contains thirty β-strands, then 27 different uninterrupted four-β-strand segments can be generated. The corresponding sub-profiles and predicted secondary structures were also tailored for the segments.

![Flowchart of WDRR](image-url)
2.3. Profile–profile alignment and candidate segment ranking

The sub-profile of each candidate segment was aligned to the template WD40 profile by an alignment script modified from our previous work (Wang et al., 2011) based on the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970). To align the position i of the candidate segment and position j of the template, the scoring function is defined as

\[ S(i, j) = w \times a_i \cdot b_j + \text{BLOSUM62}(A_i, B_j) \]  

(1)

where \( a_i \) and \( b_j \) are the vectors of position i of the candidate segment sub-profile and position j of the template WD40 profile, respectively, \( A_i \) is the residue at position i of the candidate segment, and \( B_j \) is the residue at position j of the template WD40 sequence. The dot product term and the BLOSUM62 term in Eq. (1) are combined with a weight parameter w.

The traditional affine gap penalty was employed with the opening penalty \( g_0 \) and the extension penalty \( g_1 \). Gaps were not allowed in cases where the secondary structure types of the aligning positions were both E (β-sheet). For predicted helical positions of the query segments, gaps were free to be inserted into the corresponding template positions which were not within β-strands. The gap penalty parameters \( g_0 \) and \( g_1 \), as well as the weight parameter \( w \), were trained on the training dataset (see Section 2.4) using a grid search process as described in our previous work (Wang et al., 2011). The final optimized parameters were \( w = 55 \), \( g_0 = 50 \) and \( g_1 = 3 \), respectively.

Using alignment scores between non-WD40 4-β-strand segments and the template as the background, we calculated the normalized Z-scores and p-values of candidate segments based on their raw alignment scores against the template. As shown in the flowchart (Fig. 2), all the candidate segments of the query sequence were sorted by Z-scores from the highest to lowest. Non-overlapping segments with the highest Z-scores were retained. Segments with p-value < 0.05 were chosen as predicted WD40 repeats. The query sequence would be considered as a WD40 protein only if it contained four or more predicted WD40 repeats. The boundaries of the predicted WD40 repeats were directly assigned based on the start and end positions of the alignments.

2.4. Training and testing of WDRR

We selected the sequences with less than 40% identity between each other from the all beta class in SCOP 1.75 as our training dataset. There are 2218 domains in total, 14 of which are WD40 repeat domains containing 100 WD40 repeats with the superfamiliy classification number b.69.4.x. Three non-WD40 domains were discarded due to the inability of calculating their profiles by PSI-BLAST search.

We used the original full-length protein chain sequences of “all beta” SCOP domain sequences for the testing purpose of our method. All WD40 structures in the current PDB were also included to constitute the testing dataset. These sequences were further clustered using USEARCH v4.2.66 (Edgar, 2010) at a 40% identity cutoff. Those WD40 proteins used in the training dataset were excluded from the testing dataset. In addition, we also removed those sequences that had a more than 40% identity to any sequences used for building the template. Finally, 996 sequences were obtained, 21 of which were WD40 proteins, corresponding to 163 WD40 repeats in total.

To illustrate the predictive power of the combined alignment scoring function, we compared WDRR with HMMER and HHSearch. The boundaries of each WD40 repeat were recorded from the corresponding PDB structure. The cutoff parameters (i.e. E-value or Probability) of HMMER and HHSearch were set to a very low level so that each segment could get an alignment score against the template, by which all the segments were further sorted. The performance was measured by the correct WD40 repeats found and their exact boundaries inferred from the alignments given by each method. To simplify the performance assessment, “exact boundaries” here refer to the right end only, corresponding to the WD or equivalent dipeptide position, since each WD40 repeat starts right after the preceding WD dipeptide.

3. Results and discussions

3.1. Combining dot product and BLOSUM62 leads to a better scoring function for WD40 repeat alignment and prediction

WDRR was optimized based on the training dataset. Fig. S1A shows the benchmark performance of WDRR on the training dataset, using dot product, BLOSUM62 and their combination, respectively, as the scoring functions. Dot product, as the profile–profile scoring function that considers evolutionary information, achieved a much better performance than BLOSUM62. At a very low false positive level, the performance of BLOSUM62 scoring matrix was slightly better than that of dot product. It is apparent that WDRR correctly identified more WD40 repeats using the combined scoring function than either individual component.

To illustrate why the combination of dot product and BLOSUM62 could improve the identification of WD40 repeats, we calculated the distributions of sequence identities among the top 100 segments aligned by different scoring functions, and compared with the 100 positive WD40 repeats in the training dataset (Fig. S1B–D). The sequence identities among the positive WD40 repeats in the training dataset had a wide range of approximately between 5% and 60% (solid lines). As expected, BLOSUM62 tended to align segments towards higher sequence identities, while dot product had a tendency of finding more remote homologs with low identities (dashed lines). This indicates that BLOSUM62 could not identify WD40 repeats that share remote homology, while on the other hand, dot product was too sensitive, which might result in many false positives. Thus, we combined the sequence-sequence scoring matrix (i.e. BLOSUM62) and the profile–profile scoring function (i.e. dot product) to neutralize their bias. Indeed, such expected complementation exists, as reflected by the almost identical distribution shown in Fig. S1D, based on the combined scoring function.

3.2. Performance assessment of WDRR for identifying WD40 repeats in multi-domain proteins

WD40 domains are present not only in single-domain proteins, but also frequently occur in multi-domain proteins. It is often more difficult to identify WD40 repeats from multi-domain proteins, especially the first and last repeats of the WD40 domain due to the “velcro” closure. To examine the performance of WDRR for detecting WD40 repeats from multi-domain proteins, we used the full-length sequences of the SCOP domain sequences as the test dataset (see Section 2), The performance of HMMER and HHSearch was also evaluated based on this test dataset for performance comparison.

Fig. 3A plots the true positive rates (TPR) against the false positive rates (FPR) of WD40 repeats identified by WDRR, HMMER and HHSearch. Note that the receiver-operating characteristic (ROC) curves only display the performance of up to 2% FPR. As shown in Fig. 3A, at the control of 0.01% FPR (the first 100 false positives), WDRR detected 151 WD40 repeats (92.6% TPR). That was 80% more than HMMER (83, 51.5% TPR) and 13% more than HHSearch (134, 82.2% TPR), respectively. WDRR provided a favorable performance at a more stringent false positive control: it
detected 111 true positives (68.1% TPR) before generating the first false positive. In contrast, HHsearch could only detect 14 (8.6% TPR) and HMMER did not report any true positives prior to the appearance of the first false positive.

In terms of the exact boundaries for each identified WD40 repeats, WDRR still provided a similar competitive performance, as shown in Fig. 3B, which plotted the TPR of WD40 repeats identified with the exact correct boundaries against the FPR. Note that HMMER only identified 20 true positives with correct boundaries (12.3% TPR) at 2% FPR. This is because the HMM constructed from the MSA of the WD40 repeat library contained 58 match states, which was 50% longer than the length of the WD40 repeats. It is noteworthy that the result of HMMER was strongly affected by the quality of this HMM model when being applied to find the WD40 repeat boundaries.

To deal with this problem, we directly used the seed MSA of the WD40 family (PFO0400) provided by the Pfam database (Release 25.0) to generate another template profile (used by WDRR) and the HMM (used by HMMER and HHsearch) model. This template contained 1842 WD40 repeats and the length of this HMM model was 39. The parameters of all the three methods used in this template were set as the same as the above without the need for re-training. As a result, we obtained similar ROC curves as plotted in Fig. 3C and D.

Based on the Pfam template profile, WDRR had a slight decrease in its performance of controlling false positives. But it was still able to identify 147 WD40 repeats (90.2% TPR) at 2% FPR, which was approximately 70% higher than HMMER (88, 55.8% TPR), and 7% higher than HHsearch (137, 84.0% TPR), respectively. Before reporting the first false positive, there were 44 true positives identified by WDRR (31 when considering the exact boundaries, 19.0% TPR), while HHsearch and HMMER reported only 19 (12 when considering the exact boundaries, 7.4% TPR) and 5 (3 when considering the exact boundaries, 1.8% TPR), respectively. Although the Pfam template profile contained the information from many more sequences than our structural library template profile, the results of WDRR, HMMER and HHsearch based on both profiles exhibited similar trends.

As large amounts of aligned sequences in the Pfam template helped to improve the quality of HMM, the results of HMMER and HHsearch shown in Fig. 3C and D were slightly better than those shown in Fig. 3A and B. However, WDRR performed better in the case of the structure template profile than the Pfam template profile, indicating that WDRR has a good ability to utilize useful information from structures. Besides, the Pfam template may include evolutionary information mostly from closely related WD40 segments but might lack the contribution from distantly related WD40 repeats. Irrespective of either template used, WDRR is a better method for WD40 repeat identification than HMMER and HHsearch, as strongly suggested by the results.

3.3. Comparison between WDRR and WDSP

To compare WDRR with WDSP, we run WDRR on three datasets published in the paper of WDSP. These include 33 PDB sequences that WDSP used as training set (denoted as “WDSP-training”), 68 WD40 proteins that WDSP was not able to identify (denoted as “WDSP-missed”) and 76 potential WD40 proteins that detected by WDSP (denoted as “WDSP-potential”). The results of WDSP and other methods on these three datasets were obtained directly from the supplementary material of the original WDSP paper. While WDSP recognized 3-strand ‘abc’ and 4-strand ‘abcd’ WD40 repeats, WDRR assumed all WD40 repeats to be ‘dabc’. Data file S1 provides the numbers of WD40 repeats identified in each protein by each method, regardless of the exact boundaries.

On the WDSP-training set, WDRR identified 254 WD40 repeats in the 33 PDB sequences and only missed one WD40 repeat in PDB (99.6%) with 15 false positives (6%). WDRR is much more sensitive compared to WDSP. For example, 2PM9 (PDB ID) has two chains in PDB, with eight and six WD40 repeats, respectively. The chain with eight repeats inserts one blade into the other chain. WDRR identified nine repeats in the long chain and six in the short chain, respectively, while WDSP found seven ‘abc’ repeats and only five ‘abcd’ repeats.

Of the 16 proteins in WDSP-missed, in which WDSP found too few repeats, 12 proteins have more WD40 repeats identified by WDRR. Of the 52 proteins in WDSP-missed with low WDSP score, WDRR only missed seven proteins in which no WD40 repeats were predicted, while 21 proteins have four or more identified WD40 repeats, which suggests the sensitivity of WDRR complements what WDSP would miss.

In WDSP-potential, 35 proteins have potential WD40 repeats and 36 have other types of repeats. WDRR could identify WD40 repeats from 22 out of the 35 potential WD40 proteins, in which 15 have four or more predicted WD40 repeats. WDRR identified WD40 repeats in only two proteins containing other types of repeats. Altogether, the results indicate that WDRR could more accurately predict WD40 repeats than WDSP.

3.4. Re-annotating the DUFs in the Pfam β-propeller clan

To identify those remote homologous WD40 repeats missed by Pfam, we further applied WDRR to re-annotate the WD40 Pfam family (PFO0400) as well as six DUF families out of 39 families in the β-propeller clan (CL0186). Full-length sequences of these families were downloaded from Pfam 25.0 (Finn et al., 2010). The results were further filtered by discarding proteins with fewer than four WD40 repeats identified by WDRR.

Table 1 provides the results of the identified WD40 repeats in the WD40 and three selected DUF families by WDRR (For a complete list of the re-annotation of all 39 families in the β-propeller clan, refer to Table S1 in the Supplementary material). We also
conducted WD40 repeat prediction of these Pfam families using HMMER search, providing a further comparison between WDRR and HMMER in a practical application setting.

Apparently, WDRR identified a larger number of WD40 repeats and proteins than HMMER. In the case of the WD40 family (PF00400), 14267 sequences were annotated by HMMER as WD40 proteins, containing a total of 84108 identified WD40 repeats. In contrast, WDRR identified 22344 WD40 proteins (57% more than HMMER) with a total of 179254 WD40 repeats identified (113% more than HMMER). The results suggest that on average WDRR detected two more WD40 repeats per sequence than HMMER, providing a better and larger coverage of WD40 repeats.

WDRR annotated WD40 proteins from some DUF families, which was seldom identified by HMMER. For example, WDRR annotated 60 WD40 proteins in DUF1513 and 45 in DUF2415. All the 60 proteins identified in DUF1513 represent novel WD40 proteins that were not previously annotated as WD40 proteins. There were 36 novel WD40 proteins out of the 45 proteins identified in DUF2415. In DUF1900, all the 54 WD40 proteins reported by HMMER and 221 proteins by WDRR were already included in the WD40 family. This indicates that for the WD40 proteins, a large number of distantly related WD40 repeats still could not be identified by HMMER, and were considered as a DUF domain (DUF1900) other than WD40. As a comparison, WDRR could successfully detect most of these distantly related WD40 repeats for both known and novel WD40 proteins. We further confirmed this by examining the crystal structure of the murine coronin-1A protein (UniProt AC: COR1A_MOUSE, PDB ID: 2AQ5), which contained both the WD40 and DUF1900 Pfam domains. It was also the only protein with a DUF domain in the β-propeller clan whose crystal structure has been resolved. This protein was also included in our test dataset.

As shown in Fig. 4, COR1A_MOUSE was annotated to possess one non-β-propeller domain DUF1899 (PF08953, green), three WD40 repeats (PF00400, blue), one DUF1900 domain (PF08954, magenta and yellow) and a C-terminal coiled coil (not shown in this structure) in the Pfam database. Obviously, the DUF1899 domain contains the first repeat of the WD40 β-propeller, and the DUF1900 domain contains repeats 6, 7 and the C-terminal extension against the bottom of the propeller. In contrast, HMMER could only detect three WD40 repeats (2–4, blue) of COR1A_MOUSE as in the Pfam database, while using the p-value cutoff of 0.01, WDRR could identify five repeats (i.e. repeats 2–6, blue, red and magenta) with correct boundaries. It has been discovered that the five recognizable repeats have sequence homology to the canonical WD40 repeat, but repeats 1 and 7 did not have such homology (Appleton et al., 2006). However, when we set the p-value cutoff to 0.05 (FPR < 0.35%), WDRR could correctly detect repeat 1 and partially repeat 7, while an additional false positive repeat was also reported in the C-terminal extension of DUF1900.

### 3.5. Identifying WD40 repeat proteins in model organisms

To facilitate the study of WD40 repeat proteins by the wider research community, we have implemented a web server of WDRR, and also performed proteome-wide prediction of nine model organisms. The proteomic sequences of the nine organisms were downloaded from the Ensembl project Release 60 (Flicek et al., 2011). Fig. 5 shows the statistical results of the identified WD40 repeats and proteins for each model organism. As can be seen, WD40 proteins account for about 1% of the entire proteomes, except for the percentage of 0.4% estimated for Nostoc punctiforme PCC 73102. This is consistent with the previous study of WD40 proteins (Stirmann et al., 2010). For most species studied here, 7 and 8-repeat proteins are the most abundant, followed by 4, 5, 6 and 9-repeat proteins. In Nostoc, a quite different distribution was observed, i.e. no protein with less than seven repeats was found, but approximately half of the identified WD40 proteins had more than 14 repeats. This confirmed that Nostoc is a valuable organism with a relatively high number of recently amplified WD40 propellers, making it particularly suitable for studying the evolution of WD40 proteins (Chaudhuri et al., 2008).
4. Conclusion

We have developed a novel computational method called WDRR to identify WD40 repeats based on the predicted secondary structure information and a profile–profile alignment algorithm with a combined scoring function of BLOSUM62 and dot product. The effectiveness of combining BLOSUM62 and dot product in the scoring function indicates that the identification of diverse WD40 repeats is a 'superfamily-level' task. It is difficult for traditional sequence-based search or annotation methods to identify all the repeats in one WD40 protein, especially those distantly related ones; application of methods at the fold recognition level tend to generate false positive noises more than expected. It is worth mentioning that WDRR’s great predictability benefits from the high quality of secondary structure prediction. In contrast, low quality of predicted secondary structures would lead to long candidate segments and poor global alignments, due to the big length difference between the candidate segments and the template.

In practical applications, WDRR exhibited its strength of identifying WD40 repeats and proteins with accurate assignment of the boundaries. The running speed of WDRR is lower than HMMER because it involves calculation of the sequence profiles by PSI-BLAST, but its running speed is similar to that of WDSP. However, WDRR can still be efficiently used as an alternative or extension of existing methods to obtain finer details of possible WD40 repeats and proteins. These details will not only help to accelerate functional studies of WD40 proteins without solved structures by identifying the exact positions of point mutations in the repeats, but also provide useful clues and insights into the evolution of the WD40 superfamily.

Finally, we would like to point out that the strategy proposed in this work can be generalized to predict other domains with identical secondary structural element repeats. More specifically, to achieve this, the template could be trained by using known proteins and repeats as the training set, while the candidate segments would be generated by trimming the query sequence according to the secondary structure element of the repeat. Hence, we expect that this strategy may be also exploited as a useful framework to guide and improve the prediction of other structural domains.

Acknowledgments

We thank Dr. Seung Y. Rhee at Carnegie Institution for Science for helpful discussions. This work was supported by the grants from the State Education Ministry of China (20120008110014) and the National Natural Science Foundation of China (31471249, 61202167 and 61303169). JS is a recipient of the Hundred Talents Program of the Chinese Academy of Sciences (CAS).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2016.03.025.