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

Chuan Wang[1,2], Xiaobao Dong[1,2], Lei Han[1,2], Xiao-Dong Su[3,*] and Ziding Zhang[1,2,*]

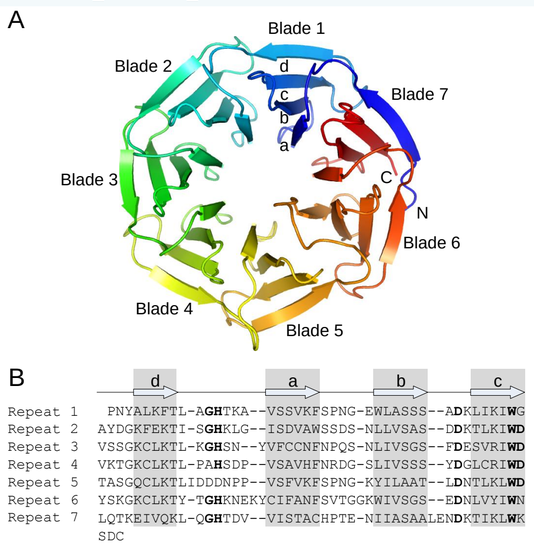
[1]State Key Laboratory of Agrobiotechnology, [2]Bioinformatics Center, College of Biological Sciences, China Agricultural University, Beijing 100193, China
[3]National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, China
[*]Corresponding authors.

Introduction

A **WD40 protein** usually 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 act as scaffolds for protein-protein interactions and are involved in many fundamental biological processes and mutations in WD40 repeats could cause diseases.

Challenges of identifying WD40 repeat proteins:

- The strand shift – "velcro" closure
- The various number of repeats in one protein
- More than one β-propeller in a single protein chain
- Short motifs or other domains inserted within WD40 repeats or between propellers



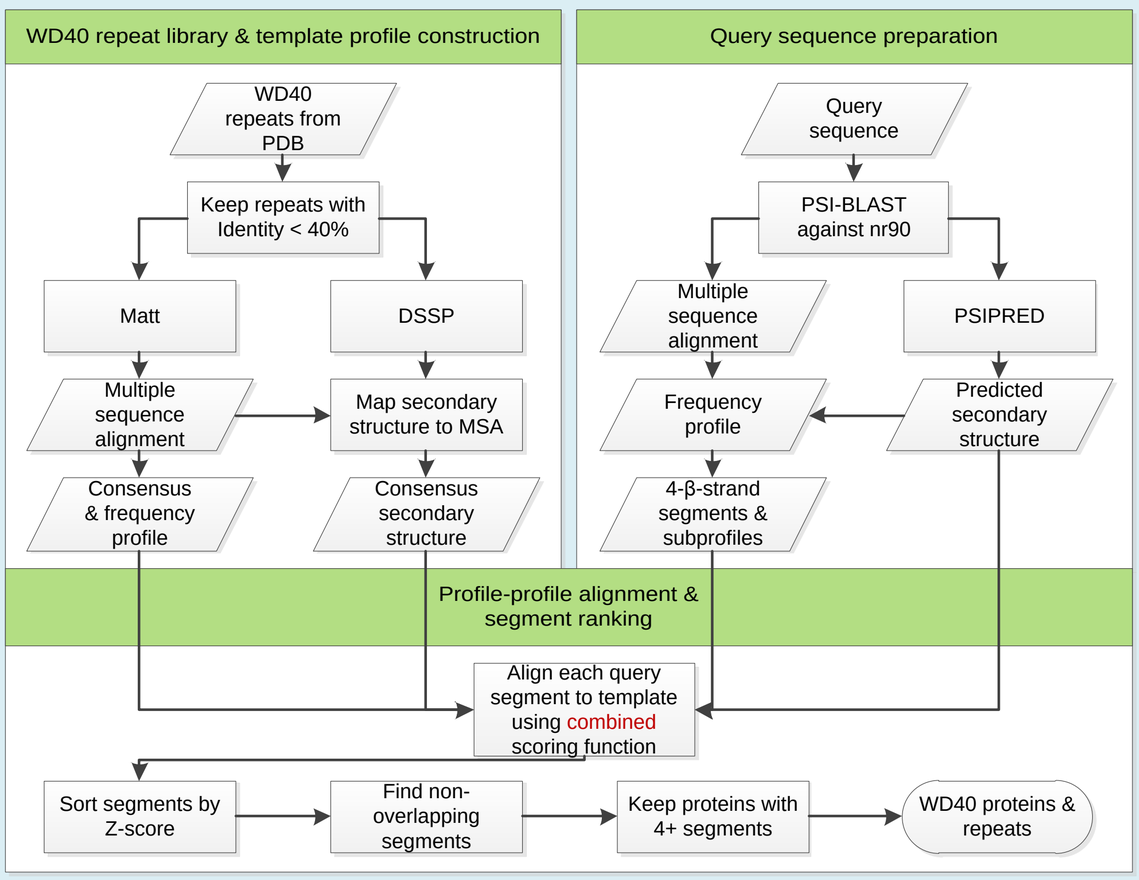
Three tiers for identifying WD40 repeat proteins:

- To determine whether a protein is a WD40 repeat containing protein
- To determine the exact number of WD40 repeats contained by the protein
- To find the exact starting and ending positions of each WD40 repeat in the protein sequence

Methods

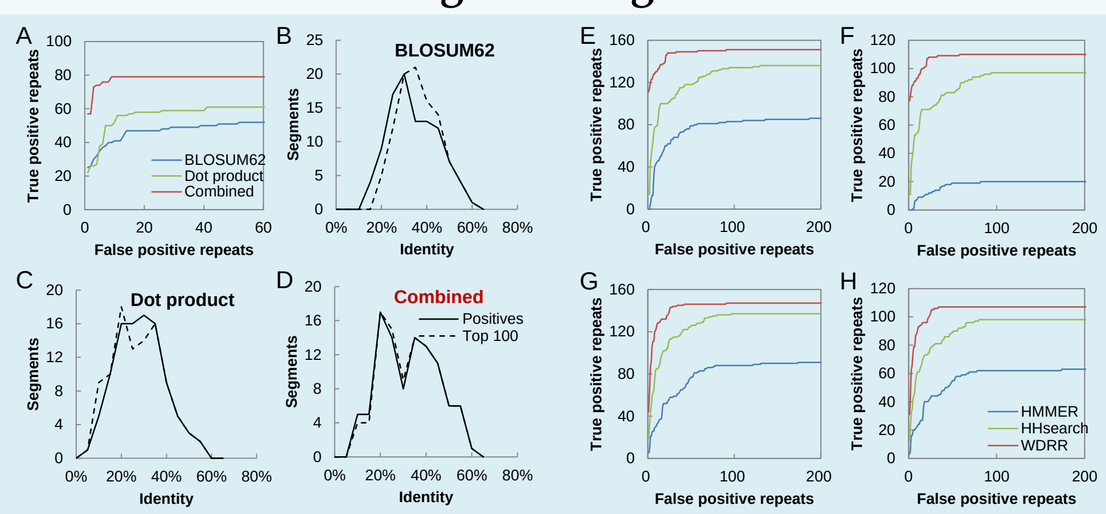
Pipeline for identifying WD40 repeats by WDRR (WD40 Repeat Recognition)

- Constructing a WD40 repeat template from WD40 structures
- Preparing query according to its predicted secondary structure
- Aligning segments using combined scoring function (BLOSUM62 & Dot product)
- Training and testing datasets were generated from the all beta class of SCOP 1.75 release and the PDB database.



Results

Performance on training & testing datasets

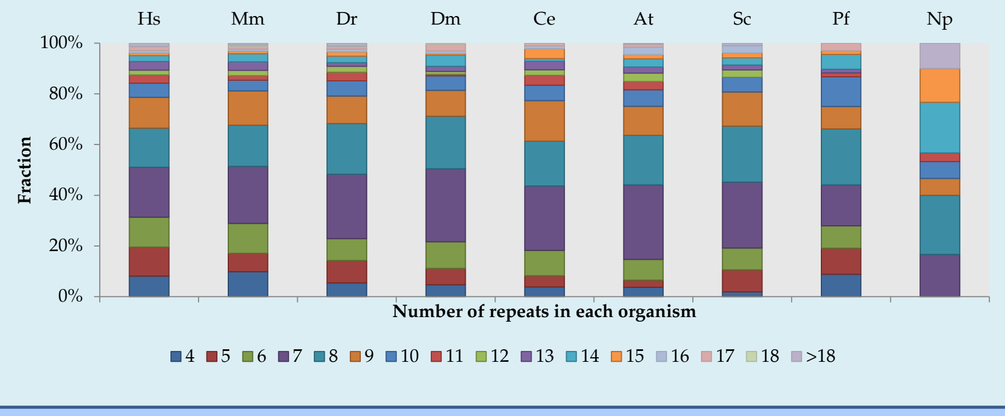


Re-annotation of the β-propeller clan (CL0186)

Pfam family	Pfam annotation			WDRR			
	#Domains	#Sequences	Avg. repeats	#Repeats	#Proteins	Avg. repeats	Novel WD proteins
WD40 (PF00400)	101999 (63.3%)	24852	4.10	199466	23180	8.61	0
Kelch_1 (PF01344)	8634 (5.4%)	3566		15	2	7.50	0
PD40 (PF07676)	7996 (5.0%)	3165		1560	174	8.97	32
RCC1 (PF00415)	7789 (4.8%)	1992		36	5	7.20	0
PQQ (PF01011)	4848 (3.0%)	1807		100	11	9.09	2
Cytochrom_D1 (PF02239)	4404 (2.7%)	3660		1156	132	8.76	21
NHL (PF01436)	3242 (2.0%)	1413		137	14	9.79	8
SGL (PF08450)	1814 (1.1%)	2221		312	28	11.14	21
Lactonase (PF10282)	1672 (1.0%)	1680		1285	134	9.59	51
Kelch_2 (PF07646)	1055 (0.7%)	2152		6	1	6.00	0
DPPIV_N (PF00930)	1012 (0.6%)	1370		41	4	10.25	2
CPSF_A (PF03178)	555 (0.3%)	536		90	12	7.50	5
eIF2A (PF08662)	529 (0.3%)	1204		8376	1020	8.21	81
Coatomer_WDAD (PF04053)	417 (0.3%)	420		4031	380	10.61	4
DUF1900 (PF08954)	343 (0.2%)	300		1922	272	7.07	0
Nucleoporin_N (PF08801)	284 (0.2%)	285		23	3	7.67	0
DUF1513 (PF07433)	173 (0.1%)	175		656	121	5.42	60
IKI3 (PF04762)	170 (0.1%)	213		1079	137	7.88	9
Nup160 (PF11715)	151 (0.1%)	280		1316	147	8.95	1
Apc4_WD40 (PF12894)	145 (0.1%)	181		714	81	8.81	1
Gmad1 (PF10647)	135 (0.1%)	284		600	85	7.06	5
Lgl_C (PF08596)	116 (0.1%)	114		988	101	9.78	40
DUF2415 (PF10313)	77 (0.0%)	74		381	62	6.15	36
Me-amine-dh_H (PF06433)	59 (0.0%)	62		7	1	7.00	0

Genome-wide identification of WD40 repeat proteins

Organisms	Hs	Mm	Dr	Dm	Ce	At	Sc	Pf	Np
#ORF	81968	50959	31473	22765	27975	27416	6696	5494	14305
#proteins									
HMMER	639	418	283	227	130	207	79	50	62
WDRR	838	568	350	295	181	245	104	68	60
#repeats									
HMMER	3825	2549	1740	1377	784	1267	470	300	720
WDRR	6844	4593	2910	2410	1527	2108	868	555	726
#average repeats									
HMMER	5.99	6.10	6.15	6.07	6.03	6.12	5.95	6.00	11.61
WDRR	8.17	8.09	8.31	8.17	8.44	8.60	8.35	8.16	12.10



Summary

- WD40 repeat identification is a superfamily level task.
- Cutting query sequences into segments according to predicted secondary structures makes the task much easier.
- WDRR effectively reduces the false positive repeats by ranking and filtering.
- Obtaining the exact boundaries of WD40 repeat may accelerate the functional and evolutional analysis of WD40 proteins.
- Web server address: http://protein.cau.edu.cn/wdrr/

Ziding Zhang's Lab of Protein Bioinformatics at China Agricultural University http://protein.cau.edu.cn
Chuan Wang's email: grittyy@cau.edu.cn