BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

BIOMEDICAL INFORMATICS

www.bioinformation.net

Volume 12(5)

Hypothesis

A Comparison of Rosetta Stones in Adapter Protein Families

Hulikal Shivashankara Santosh Kumar[1], Vadlapudi Kumar[2*]

[1]Department of Biotechnology and Bioinformatics, Kuvempu University, Shankaraghatta - 577451, Karnataka, India - E-mail: sk.genesan@gmail.com; [2]Department of Biochemistry, Davanagere University, Shivagangothri, Davanagere - 577002, Karnataka, India - Email: vadlapudikumar@gmail.com; * Corresponding author

Received July 12, 2016, Revised July 24, 2016; Accepted July 26, 2016; Published August 15, 2016

Abstract:
The inventory of proteins used in different kingdoms appears surprisingly similar in all sequenced eukaryotic genome. Protein domains represent the basic evolutionary units that form proteins. Domain duplication and shuffling by recombination are probably the most important forces driving protein evolution and hence the complexity of the proteome. While the duplication of whole genes as well as domain encoding exons increases the abundance of domains in the proteome, domain shuffling increases versatility, i.e. the number of distinct contexts in which a domain can occur. In this study we considered five important adapter domain families namely WD40, KELCH, Ankyrin, PDZ and Pleckstrin Homology (PH domain) family for the comparison of Domain versatility, Abundance and domain sharing between them. We used ecological statistics methods such as Jaccard's Similarity Index (JSI), Detrended Correspondence Analysis, k-Means clustering for the domain distribution data. We found high propensity of domain sharing between PH and PDZ. We found higher abundance of only few selected domains in PH, PDZ, ANK and KELCH families. We also found WD40 family with high versatility and less redundant domain occurrence, with less domain sharing. Hence, the assignments of functions to more orphan WD40 proteins that will help in the identification of suitable drug targets.

Keywords: WD40, Rosetta Stones, domain versatility, detrended correspondence analysis, domain sharing

Background:
The building blocks that create protein three dimensional structures are called domains, and domains are often combined to create multi-domain proteins or tethered proteins and the process is called as "domain tethering". In many vertebrate proteins, repeats with several adjacent domains from the same family can be found **[1].** During evolution, they have been duplicated, fused and recombined, to produce proteins with novel structures and functions **[2].** Comparisons of the proteomes of different organisms have suggested that proteins have evolved increasingly complex functions primarily by the acquisition of pre-existing domains, resulting in the formation of new multi domain architectures, whereas the emergence of an entirely new domain is a relatively rare event **[3].**

Domains can recombine to form multi-domain proteins, and proteins with two or more domains constitute the majority of proteins in all genomes. Thus, the recombination of existing domains may be a major mechanism that modifies protein function and increases proteome complexity **[3].** The combination or shuffling of domains increases what is termed as the versatility of a domain superfamily; that is, the number of different partner domains that domains of a particular superfamily are adjacent to. The extent of duplication of different combinations varies widely and, in nature, will depend on selection for the domain combination based on its function. Some of the pair-wise domain combinations that are highly duplicated also recur frequently with other partner domains **[4].**

We studied five families having namely WD40, ANKYRIN, KELCH, PDZ and PH. The study pool contained WD40 and Ankyrin repeat (ANK) families which are among the most frequently occurring repeats among eukaryotes **[1].** KELCH and WD40 repeats consist of repeated sequence motifs with hallmark residues spaced at regular intervals. Significant diversity has been observed in both WD40 and KELCH repeat sequences, a large number have repeat lengths and repeat spacing, yet they resemble same three dimensional Beta propeller structures **[5].** PDZ family is considered because they are abundant protein interaction modules that often recognize short amino acid motifs

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

at the C-termini of target proteins and are known to regulate many signalling pathways **[6].** Pleckstrin Homology (PH) domains have been known to have multiple roles but predominantly involved in Inositol phosphate signalling **[7].** All the above families contain protein with high degree interaction partners indicating their participation in interactome **[8].**

ANK belongs to all alpha class, PDZ and PH belongs to alpha-beta class which are predominantly found at the cell membrane, whereas KELCH and WD40 belong to all beta class of proteins. WD40 and KELCH have been recognised for their roles in transcription regulation and protein ubiquitination respectively. ANK plays role in various functions from signal transduction to transcriptional regulation. ANK form folded solenoid structure **(Figure 1a),** PH domain is made of two perpendicular anti-parallel beta sheets, followed by a C-terminal amphipathic helix, WD40 and KELCH proteins form propeller structure.But none of the above mentioned domains have inherent catalytic activity and they act as modules for protein-protein interactions thereby regulating the process in which they are involved. The structural and functional diversity exhibited by the families is going to reflect in the sample sequences too. Hence, the study pool spans topologically and functionally diverse proteins. We compared the domain tethering pattern in each family, domain sharing between the families and by applying methods for ecological data analysis first we analysed the similarity and then we analyse the difference in terms of the domain composition. We tried to identify most diverse and most skewed families there by finding the unique domain family among those taken in this study.

Table 1: Jaccard's similarity index matrix. JSI values are high for PH, PDZ and ANK whereas least for WD40 and KELCH

	WD40	PDZ	PH	KELCH	ANK
WD40	--	--	--	--	--
PDZ	0.038	--	--	--	--
PH	0.068	0.194	--	--	--
KELCH	0.024	0	0	--	--
ANK	0.076	0.079	0.113	0.028	--

Methodology:

Construction of local repository:

The sequences representing the five families were retrieved using profile HMMs at Pfam database according to method prescribed by Krupa and Srinivasan **[19].** Each sequence was manually curated for the gene name using NCBI GENE and Uniprot database. Sequence redundancy was removedmanually, domain repertoire were catalogued in Excel file and formatted to DB format using PERL program. Front end of database was created using WAMPP architecture with HTML-PHP script as front end and MySQL as back end with PERL as query program.

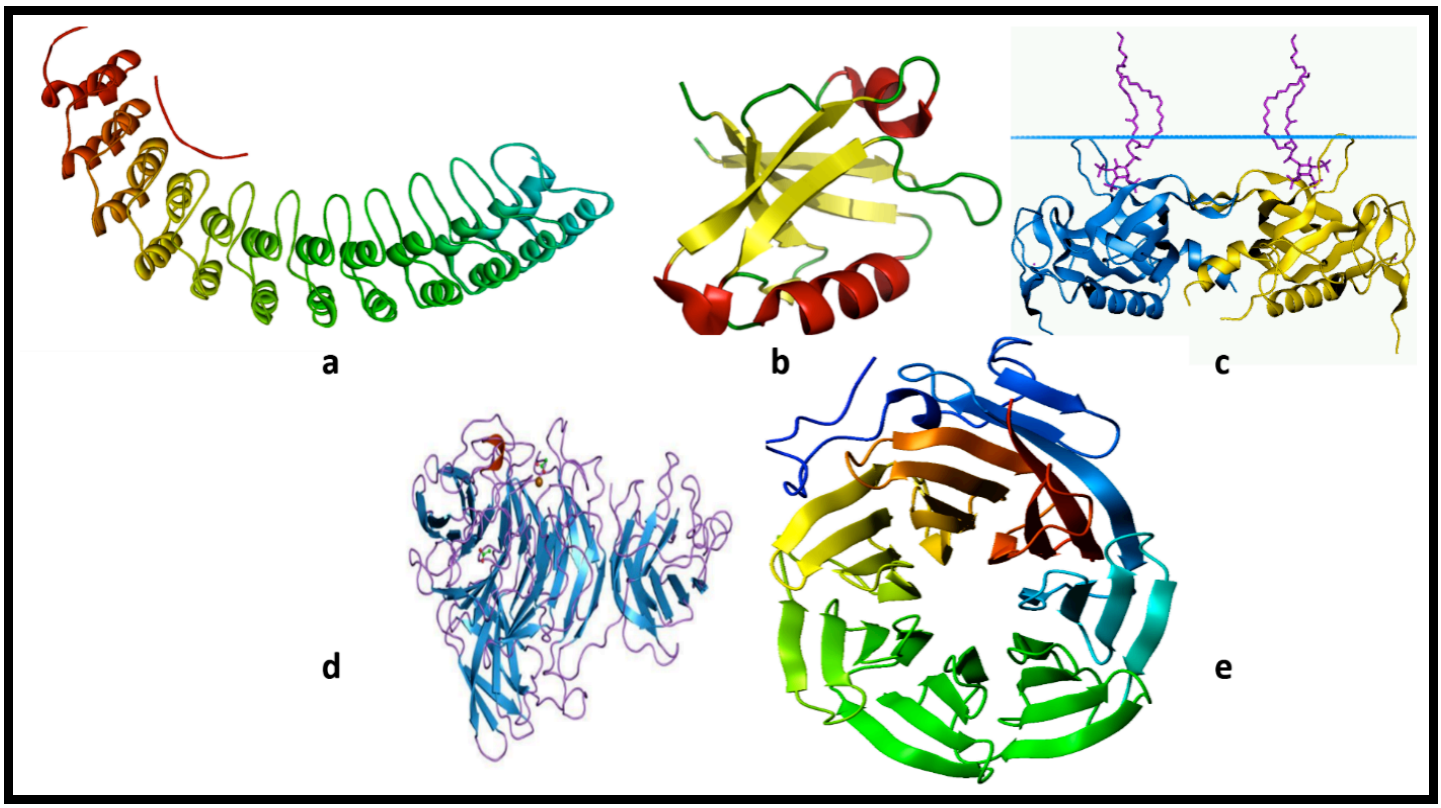

Figure 1: Representative structures of the Domains studied. a. Ankyrin repeat, b. PDZ domain, c. PH domain, d. KELCH repeat and e. WD40 repeat structure.

BIOMEDICAL INFORMATICS

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

Counting the domain tethering events:
Domain tethering analysis was done according to Krupa and Srinivasan **[19]** Domain A is found covalently attached with Domain B in a protein sequence. With reference to Domain A, Domain B is considered as tethered domain with count 1. In the protein 2 if Domain A is found covalent attached with Domain B and Domain C, the tethering number is 2. The frequency of domains occurring among the above said five families were also recorded.

Data analysis:
A matrix was designed taking families as columns and domains as rows. For similarity analysis, a matrix was created with domains in rows and families in columns giving a score 1 for presence of domain and 0 for absence of domain in particular family under consideration. Jaccard's similarity index (JSI) was calculated based on method proposed by Real and Vargas, 1996 **[9].** A Cluster analysis was done using Euclidian distances and Ward's method.

Another matrix was created indicating frequency of domain occurrences in respective families **(Table 2).** Detrended Correspondence Analysis (DCA) of the matrix was done to identify the unique family. All the statistical analysis was done using PAST programme **[22].**

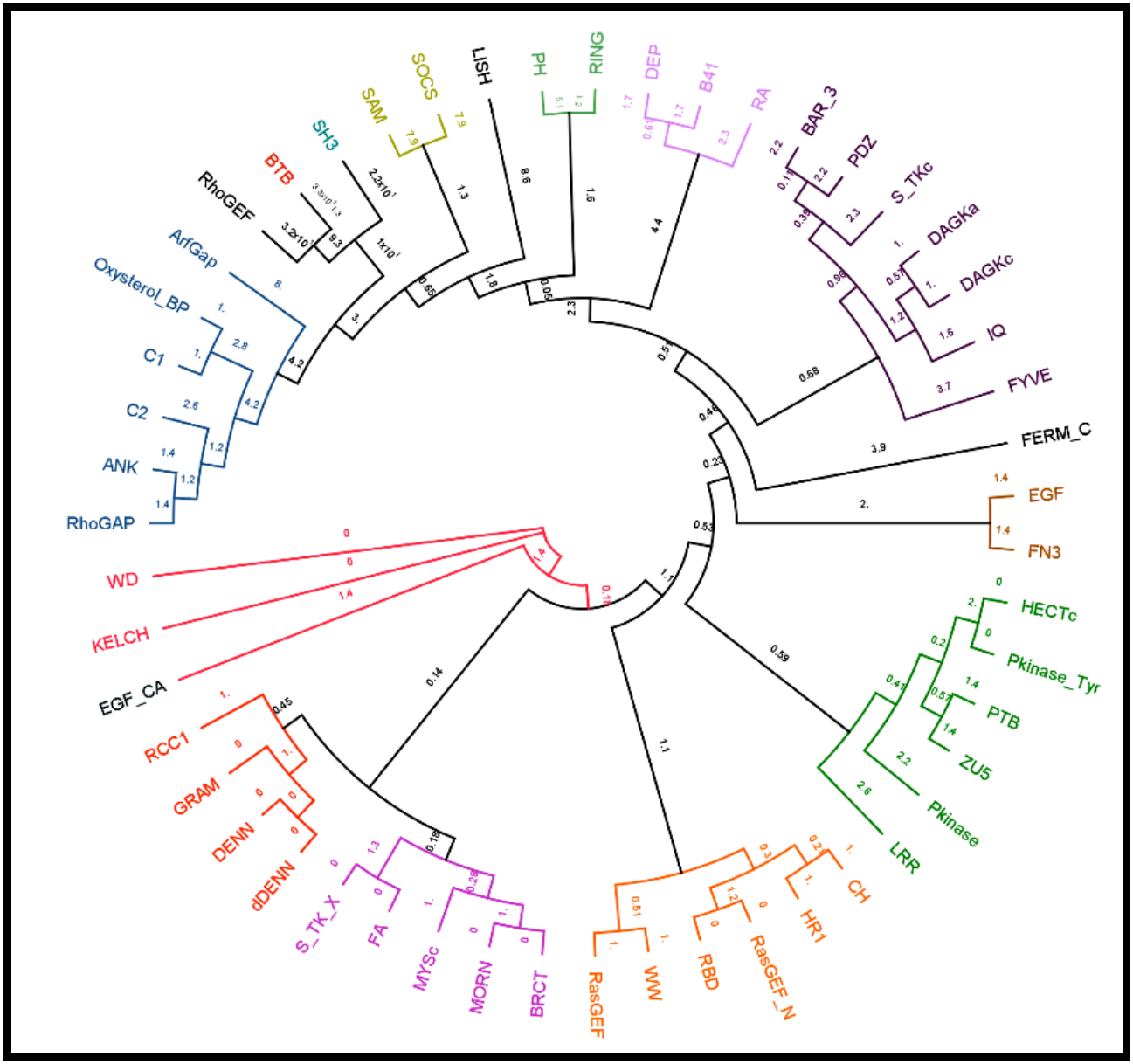

Figure 2: JSI based clustering of tethered domains (only significant domains shown).

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

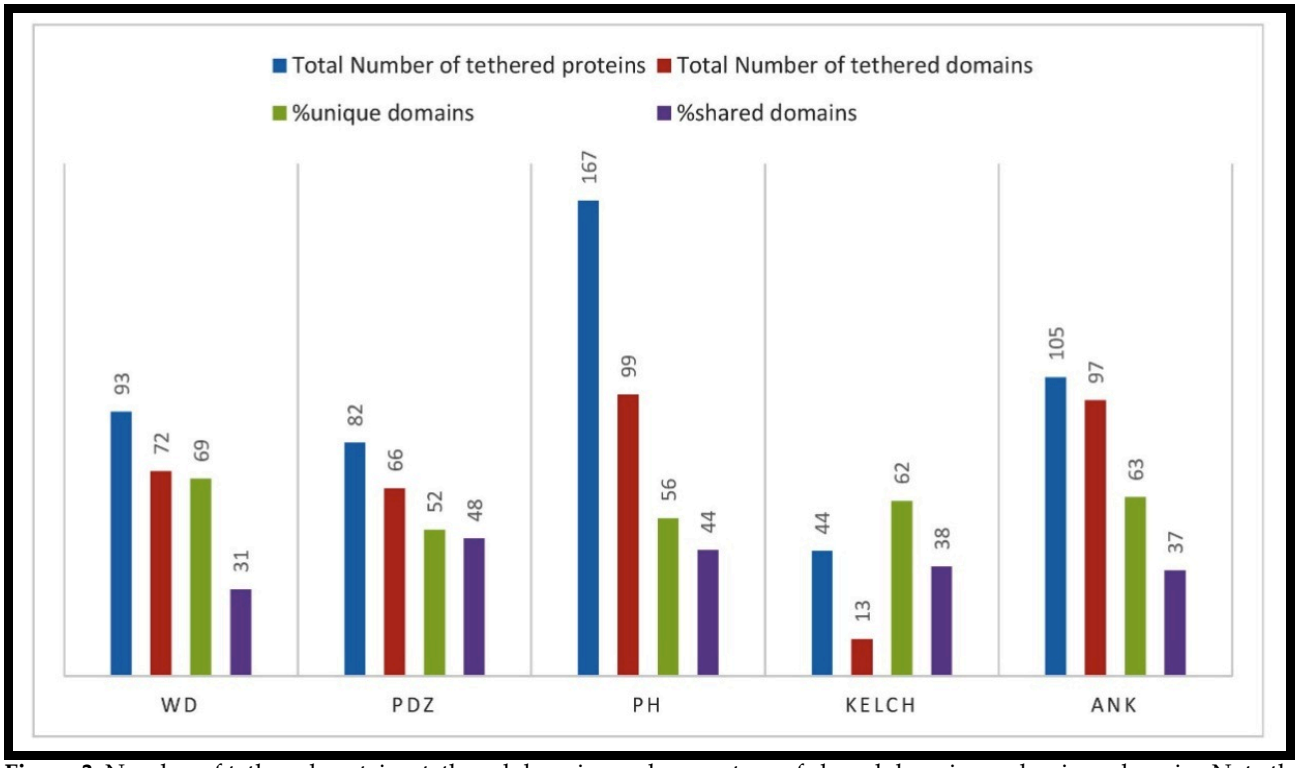

Figure 3: Number of tethered proteins, tethered domains and percentage of shared domains and unique domains.Note that PH is highest in tethering number while KELCH has least tethering number, WD has highest percentage of unique domains.

Table 2: Domain composition profile of different families under study

S. No	Domains	WD	PDZ	PH	KELCH	ANK	S. No	Domains	WD	PDZ	PH	KELCH	ANK
1	ANK	1	1	13	0	0	26	LISH	8	0	0	1	0
2	ArfGap	0	0	18	0	6	27	LRR	2	2	1	0	2
3	B41	0	7	5	0	1	28	MORN	0	0	1	0	1
4	BAR_3	0	0	6	0	2	29	MYSc	0	1	1	0	1
5	BRCT	0	0	1	0	1	30	Oxysterol_BP	0	0	11	0	1
6	BTB	2	0	0	32	3	31	PDZ	0	0	5	0	0
7	C1	0	0	11	0	2	32	PH	0	7	0	0	5
8	C2	0	3	15	0	0	33	Pkinase	1	0	0	0	4
9	CH	0	1	3	0	0	34	Pkinase_Tyr	1	0	0	0	2
10	DAGKa	0	0	3	0	2	35	PTB	0	2	0	0	2
11	DAGKc	0	0	4	0	2	36	RA	0	5	6	0	0
12	dDENN	1	0	1	0	0	37	RasGEF	0	2	3	0	0
13	DENN	1	0	1	0	0	38	RasGEF_N	0	2	2	0	0
14	DEP	0	6	4	0	0	39	RBD	0	2	2	0	0
15	EGF	0	0	0	3	1	40	RCC1	1	0	1	0	1
16	EGF_CA	0	0	0	1	1	41	RhoGAP	0	2	13	0	0
17	FA	0	1	1	0	0	42	RhoGEF	0	3	35	0	0
18	FERM_C	0	4	1	0	0	43	RING	3	3	0	0	4
19	FN3	0	0	0	2	2	44	S_TK_X	0	1	1	0	0
20	FYVE	3	0	6	0	1	45	S_TKc	0	2	5	0	1
21	GRAM	1	0	1	0	0	46	SAM	1	3	7	0	8
22	HECTc	1	0	0	0	2	47	SH3	1	17	18	0	6
23	HR1	0	1	2	0	0	48	SOCS box	3	0	0	0	9
24	IQ	1	1	3	0	2	49	WD	0	0	0	0	0
25	KELCH	0	0	0	0	0	50	WW	0	2	4	0	0
							51	ZU5	0	1	0	0	3

Note: Except WD40 family, all the others has redundant occurrence of few selected domains

BIOMEDICAL INFORMATICS

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

Results and Discussion:

Domains impart the structure and function to a protein. Due to exon shuffling and recombination many domains tend to occur in a single polypeptide called tethered proteins also called as Rosetta stone proteinwhich are the hallmarks of protein evolution **[10].** The analysis of domain tethering has been described in methods section. The total number of tethered protein and total number of tethered domain is shown in **Figure 3.** It is evident that, though the number of tethered proteins is high, the tethering number is low in case of PH and KELCH. The ratio of tethered protein to tethered domain was found to be 0.65, 0.63, 0.23, 0.11 and 0.66 for WD40, PDZ, PH, KELCH and ANK families respectively. It indicates that, PH and KELCH may have the repetitive combination of same domains that may give paralogs with different functions. There may be other situation involving too many isoforms of the same protein with different tissue specific expressions. In such a case due to the repetitive occurrence of same domains, the function of the family also will be skewed towards few biological processes.

Protein domains either may be found shared between different protein families or will be strictly confined to one particular family in some cases. All the families under consideration have shown both situations **(Figure 3).** WD40 family shows highest percentage of unique domains indicating most of the domains occurring in WD40 family are confined to itself, we call them as unique domains for convenience. However, on the other hand PDZ domain has highest percentage of shared domains. This domain distribution is similar to classical biodiversity data analysis where the similarity and dissimilarity is examined to decide the extent of diversity of flora and fauna of particular area under consideration for which JSI has been used widely. JSI provides the association between different entities in a data distribution. The scores in the matrix represent the level of association between the families under consideration. Hence higher the score higher the number of domains shared between two families. JSI provides the association between different entities in a data distribution. It has been used earlier in various cases such as clustering protein similarity networks **[12]** domain architecture comparison for multidomain homology **[13]** and automatic classification of protein structures **[14].**

There is a greater association between PDZ and PH families followed by PH and ANK. However, PH and PDZ families did not found to be sharing domains with KELCH. The maximum score for KELCH and other family association is 0.028 indicating that KELCH has limited domain sharing. WD40 domain has maximum value of 0.076 with ANK and minimum of 0.024 with KELCH. The highest level of domain association is between PH and PDZ followed by PH and ANK **(Table 1).** It is clear that there is certain type of propensity towards domain distribution between the families. When the domains were clustered using JSI score **(Figure 2)** WD40 and KELCH family were clustered as separate out groups substantiating the observations in the JSI matrix. This hints the uniqueness of WD40 and KELCH families with limited domain sharing tendencies.

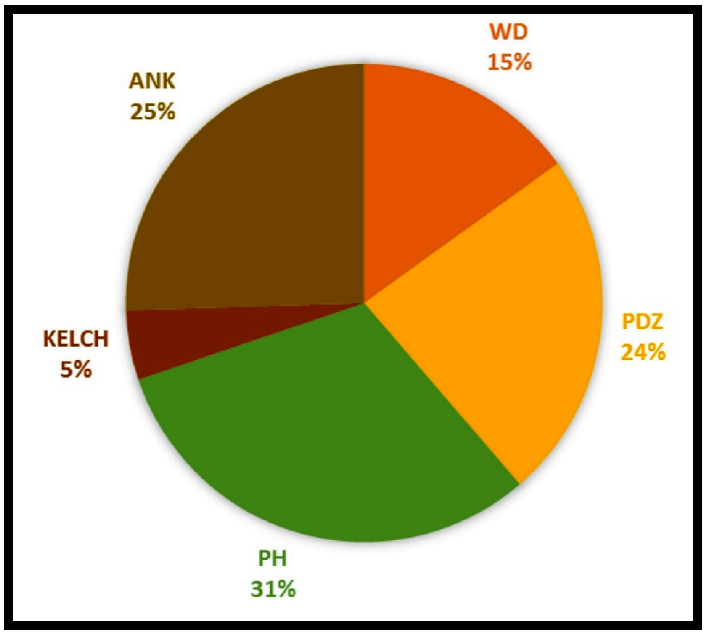

Figure 4: Contribution of domain for sharing by different families. Note that KELCH and WD40 contribute less towards domain sharing.

Domain distribution pattern were later observed for the shared domains along with their frequencies of occurrences across the five families considered in this study (Table. 2). A detrended correspondence analysis which is a part of community ordinance method was done for the dataset. Detrended correspondence analysis (DCA) is a multivariate statistical technique to find the main factors or gradients in large data. DCA is an iterative algorithm that has shown it to be a highly reliable and useful tool for data exploration and summary in community ecology **[15].** Observations from the DCA plot **(Figure 5),** domain sharing data (Figure 4), Jaccard's similarity index table **(Table 1)** and domain sharing table **(Table 2)** makes it clear that WD40 and KELCH are unique families. Out of 51 shared domains KELCH contributes to 5% and WD40 contributes 16 out of 51 shared domain which is 15% of the total domain repertoire listed in table 2 **(Figure 4).** Together KELCH and WD40 contribute to only 20% to domain sharing across different families.

The highlighted cells in the table **(Table 2)** represent the frequency of occurrence of domains shows LISH domain is the only domain with maximum frequency of 8 to occur in WD40 family. BTB dominates KELCH, SH3 and B41 dominates PDZ, RhoGEF and SH3 dominates PH and ArfGAP, SAM and SOCSboxdomains dominates ANK in composition. Since the domain composition imparts the function through providing specific geometry to the 3D structure of proteins, it can be hypothesised that, skewness in the domain composition leads to skewness in the protein function also **[16].**

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

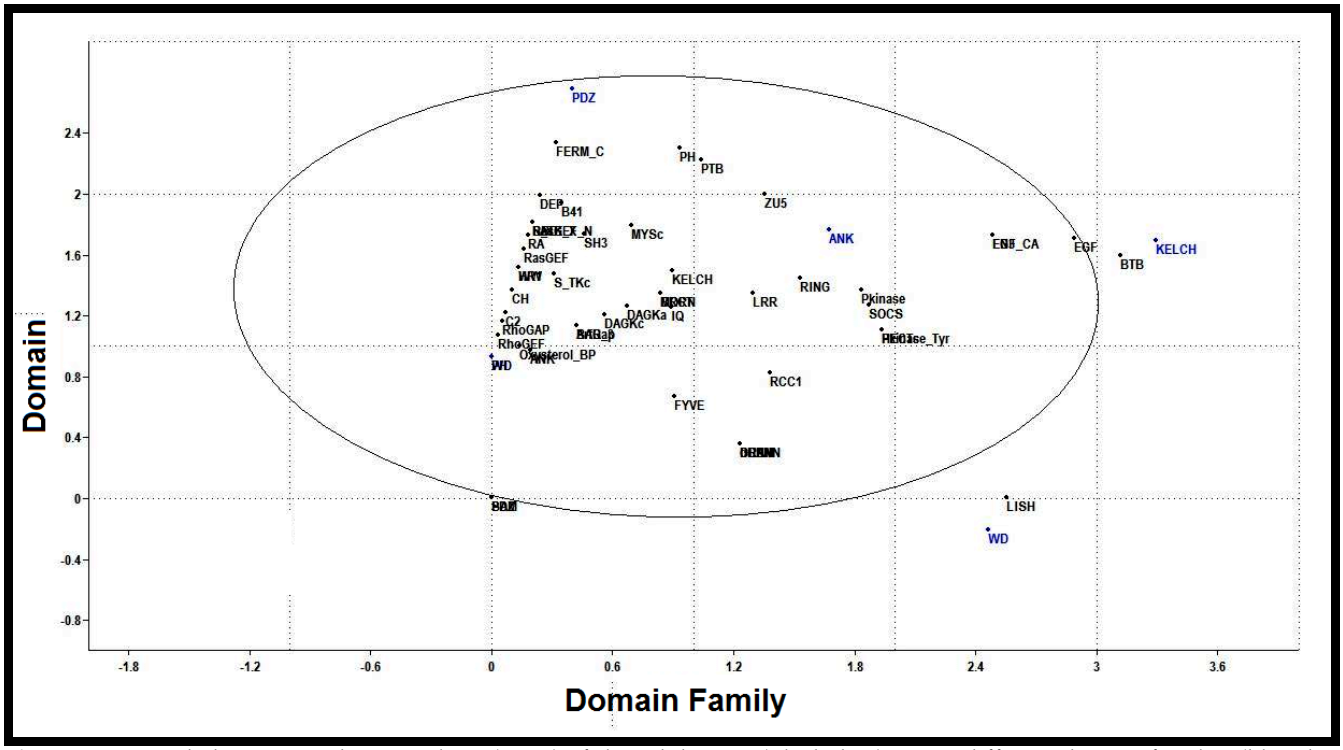

Figure 5: Detrended correspondence analysis (DCA) of shared domain (Black dots) across different domain families (blue dots). Note that WD40 and KELCH lies outside signifying their uniqueness.

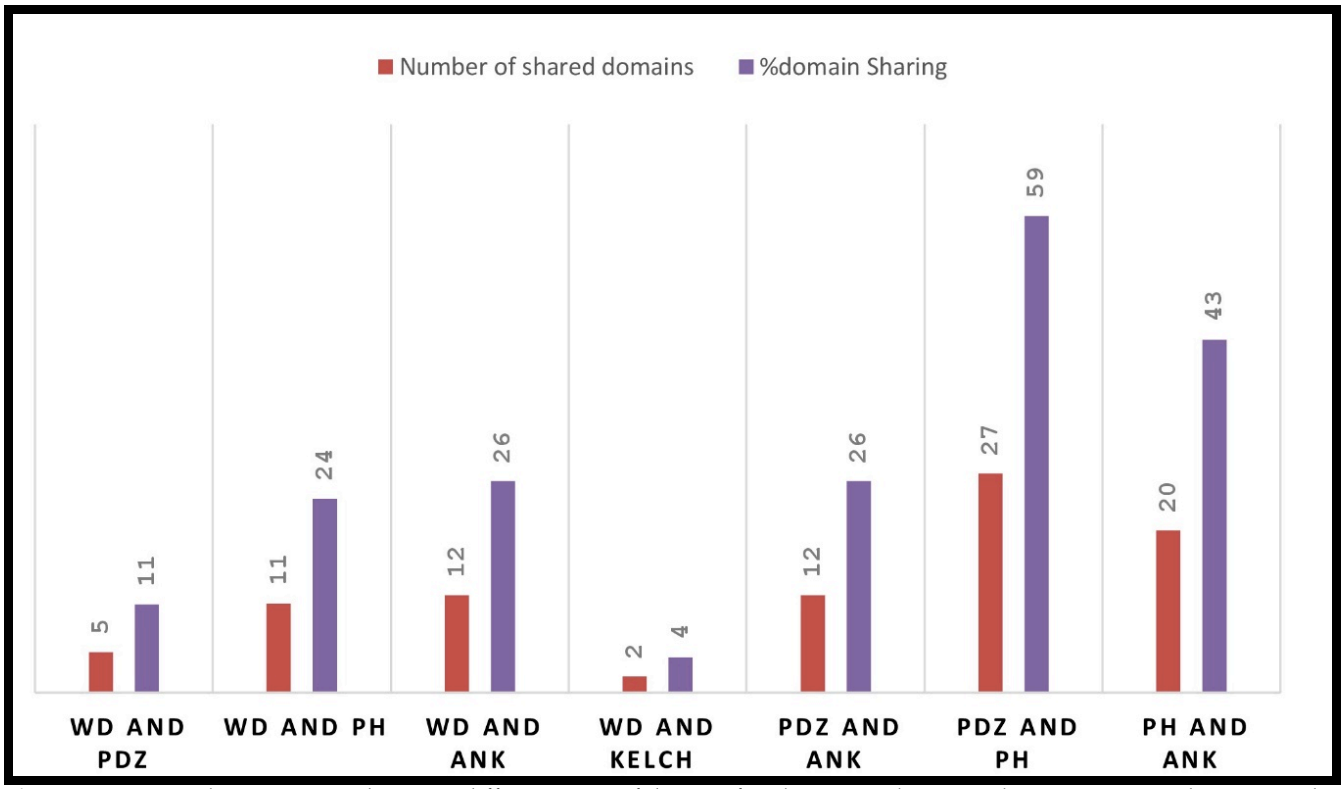

Figure 6: Domain sharing pattern between different pairs of domain families. Note that PDZ shares maximum domains and KELCH shares minimum number of domains.

ISSN 0973-2063 (online) 0973-8894 (print)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

Table 3: Domain sharing between different families

S. No.	PDZ and ANK	WD40 and ANK	WD40 and PDZ	PH and PDZ	PH and ANK	WD40 and PH	KELCH and ANK	KELCH and WD40	PH and KELCH	PDZ and KELCH
1	ANK	BTB	IQ	ANK	ANK	DENN	BTB	BTB	0	0
2	B41	COR	LRR	B41	ArfGap	FYVE	EGF	LISH		
3	IQ	FYVE	RING	C2	B41	GRAM	FN3			
4	LRR	HECTc	SAM	CH	BAR_3	IQ				
5	MYSc	IQ	SH3	CRIC_ras_sig	BRCT	KISc				
6	PH	LRR		DEP	C1	LRR				
7	PTB	PKINASE		DUF1170	DAGKa	RCC1				
8	RING	Pkinase_Tyr		FA	DAGKc	RCC1_2				
9	SAM	RCC1		FERM_C	FN3	SAM				
10	SH3	RING		FHA	FYVE	SH3				
11	S_TKc	SAM		IQ	IQ	dDENN				
12	ZU5	SH3		LRR	LRR					
13		SOCS Box		MYSc	MORN					
14		TPR		PDZ	MYSc					
15		ZnF_C2H2		PH	Oxysterol_BP					
16				PX	PH					
17				RA	RCC1					
18				RBD	SAM					
19				RGS	SH3					
20				RasGEF	S_TKc					
21				RhoGAP						
22				RhoGEF						
23				SAM						
24				SH3						
25				S_TK_X						
26				S_TKc						
27				WW						

Note: PH and PDZ have the most number of common domains.

WD40 family has high versatility because it comprises highest percentage of unique domains **(Figure 3).** It is evident that WD40 is not interested in sharing common domains with other families and there exists less domain redundancy unlike KELCH **(Figure 4 and Table 3).** KELCH has a high abundance of BTB domain but has lesser versatility suggesting the functional skewness **(Table 2).** Hence it can be concluded that, KELCH is skewed towards specific function whereas WD40 perform different function but with set of domains dedicated to only WD40 family. Higher domain shuffling results in higher number of tethering number which imparts functional versatility **[11]** to the domain family lacking which the versatility is lost. Hence, it may result in skewness towards certain biological functions in the family. It is clear that some domains are shared specifically between two families and sometimes they are restricted to single family. Observing all the above data and behaviour of the families, it makes clear that WD40 family is unique with respect to domain tethering behaviour.

PDZ and PH families were found to share their domains (59%) more often followed by PH and ANK (43%) **(Table 3)** with JSI 0.194 and 0.113 respectively **(Table 1),** which is the maximum score in the table. This indicates the extensive domain sharing between two families. On the contrary, KELCH and ANK, KELCH and WD40 have least amount of domain shared with 7% and 4% respectively and their JSI score is 0.028 and 0.024 respectively. This may be due to very less tethered domains **(Figure 3)** and lesser shared domains (Table 3 and Figure 4).Also, the domain sharing in a pair involving WD40 domain is less (Fig. 6). For example, the domain sharing is 11% (WD40 and PDZ), 24% (WD40 and PH) and 26 % (WD40 and ANK) with JSI 0.038, 0.068 and 0.076 respectively. Unlike KELCH, WD40 family despite of having six fold higher domain repertoires, WD40 still have very less contribution for domain sharing accounting to only 15% **(Figure 4).** This shows there is versatility in the domain repertoire of WD40 family but most of them are restricted to WD40 family only.

WD40 has been one of the top 10 most promiscuous domains in eukaryotes. WD40 has been reported to be one of the highly connected, and therefore likely have multiple potential functions and would not be restricted to any particular functional branch. WD40, since it has high versatility, it is regarded to be among top 10 highly social domain club meaning, the larger set of clubs contains proteins with multiple distinct domains **[3].** The nature always has preferred WD40 domain because they pose greater symmetry in structure in contrast to other abundant domains that predominate in intra- cellular processes. The reason for the

symmetry is regular repeating super-secondary structure elements. The beta propeller scaffold always allows long insertions or deletions or multiple single amino acid substitutions evolving a new binding site for an interaction partner. In WD40 propeller scaffold, there is no interlocking of secondary structures unlike TIM barrel domains which allows mutations to occur without drastic effect on structure of the scaffold. PDZ, PH, SH3 are complex structures in which there is no such system of regular repeats, rendering drastic changes in the protein sequence more likely to disrupt the overall structure **[17].** In isolation, WD40 domains have posed challenge to characterize and study, probably because they are often subunits in larger assemblies, but also because, in most cases, they lack measurable intrinsic catalysis. Whatever the reason for the adaptability of WD40 domains to act as scaffolds, they clearly represent one of the most important domain families for most critical cell processes **[18].**

Many earlier studies have implicated the WD40 proteins in various ailments mostly related to cancer and other developmental disorders **[19].** Several reports have shown that WD40 is one of the highly social family meaning, the members of the family are having higher number of interacting partners **[20].** Recently WD40 proteins have also been demonstrated to have roles in lifestyle borne diseases like hyperlipidaemia, diabetes **[21]** and so on.

This emphasizes that, due to domain structure and composition diversity, WD40 proteins are able to interact with various proteins making high degree protein interaction network, regulating various different biological processes yet there are limited publications on members of this family. Many WD40 proteins are still largely regarded as WD Repeat (WDRs) containing proteins only. WDRs have not been clearly deduced with their gene ontology even in any knowledge bases. A deeper understanding of their structures, interactions and functional diversity will be crucialfor our understanding of detailed cellular processes, and ultimately might provide new means to tinker with biological functions via synthetic and systems biology approacheswhich in turn may open new avenue for identifying new potential biomarkers.

Conclusion:

In the present study, we have compared five adapter protein families with respect to their domain composition and domain distribution among them. We found ANK, PH and PDZ families share their domain more often than WD40 and KELCH. We applied ecological tools to domain distribution data. The analysis has helped us to find how unique the familes with respect to domain composition. We found high degree of redundancy with respect to domain composition in all families except WD40 and we also found WD40 with highest percentage of versatility. In the light of the fact indicating limited study on WD40 protein evident by limited number of publications in public domain, we propose to give more emphasis to this family which helps deorphanizing the unknown proteins of the family.

Conflict of Interest

Authors declare that there is no conflict of interest.

Acknowledgement:

Authors would like to thank Prof. S. V. Krishnamurthy, Department of Environmental Science, Kuvempu University, Nikethan Lancer D'Souza, and Mrs. Prathibha JS for lending their valuable inputs in designing the database and data analysis. Prof. Riaz Mahmood, Department of Biotechnology, Kuvempu University and Ms. Rizwana Abid for her valuable inputs during manuscript preparation.

References:

[1] Björklund ÅK *et al*. *PLoSComput Biol*. 2006 **2(8)**: 0959. [PMID: 16933986].
[2] Kinch LN and Grishin N V. *CurrOpinStruct Biol*. 2002 **12(3)**: 400. [PMID: 12127461].
[3] Jin J et al. *Sci Signal*. 2009 **2(98)**. [PMID: 19934434].
[4] Vogel C *et al*. *J Mol Biol*. 2005 **346(1)**: 355. [PMID: 15663950].
[5] Clemen CS *et al*. *SubcellBiochem*. 2008 **48**:1.
[6] Lee H-J and Zheng JJ. *Cell Commun Signal*. 2010 **8**:8. [PMID: 20509869].
[7] Saraste M and Hyviinen M. *Curr Opin Struct Biol*. 1995 **403**: 8. [PMID: 7583640].
[8] Pawson T *et al*. *FEBS Lett*. 2002; **513(1)**: 2. [PMID: 11911873].
[9] Real R and Vargas JM. *SystBiol*. 1996 **45(3)**: 380.
[10] Bornberg-Bauer E*et al*. *Cell Mol Life Sci*. 2005 **62(4)**:435. [PMID: 15719170].
[11] Apic G *et al*. *Bioinformatics*. 2001 **17 (1)**: S83. [PMID: 11472996].
[12] Mai T-L *et al*. *J Proteome Res*. 2016 **15(7)**: 2123. [PMID: 27267620].
[13] Song N *et al*. *J Comput Biol*. 2007 **14(4)**: 496. [PMID: 17572026].
[14] Santini G *et al*. *BMC Bioinformatics*. 2012 **13(1)**: 233. [PMID: 22974051].
[15] Shaw PJA. 2003; Hodder-Arnold, London
[16] Ponting CP and Russell RR. *Annu Rev BiophysBiomolStruct*. 2002 **31**: 45. [PMID: 11988462].
[17] Chaudhuri I *et al*. *StructFunct Genet*. 2008 **71**: 795. [PMID: 17979191]
[18] Stirnimann CU *et al*. *Trends. Biochem. Sci*. 2010 **35**:565. [PMID: 20451393]
[19] Li D & Roberts R. Cell Mol Life Sci. 2001 58(14): 2085. [PMID: 11814058].
[20] Krupa A & Srinivasan N. *Gene*. 2006 **380**: 1. [PMID: 16843620]
[21] Hulikal Shivashankara SK *et al*. *Bioinformation*. 2016 **12(2):** 54
[22] Hammer O *et al*. *Palaeont. Elect*. 2001 **4(1)**: 9

Edited by P Kangueane

Citation: Santosh Kumar & Kumar, Bioinformation 2016 12(5) 285-292.