BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net ***Hypothesis***

Volume 9(10)

An alignment-free domain architecture similarity search (ADASS) algorithm for inferring homology between multi-domain proteins

Divya P Syamaladevi[1, 2], Adwait Joshi[2] & Ramanathan Sowdhamini[2*]

[1]Sugarcane Breeding Institute Indian Council of Agricultural Research Coimbatore, India, PIN 641 007; [2]National Center for Biological Sciences (TIFR), UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India; Ramanathan Sowdhamini– Email: mini@ncbs.res.in; Phone: +91-080-23666001; FAX: +91-080-23636421; *Corresponding author

Received November 19, 2012; Revised January 01, 2013; Accepted January 02, 2013; Published June 08, 2013

Abstract:
Annotations of the genes and their products are largely guided by inferring homology. Sequence similarity is the primary measure used for annotation purpose however, the domain content and order were given less importance albeit the fact that domain insertion, deletion, positional changes can bring in functional varieties. Of late, several methods developed quantify domain architecture similarity depending on alignments of their sequences and are focused on only homologous proteins. We present an alignment-free domain architecture-similarity search (ADASS) algorithm that identifies proteins that share very poor sequence similarity yet having similar domain architectures. We introduce a "singlet matching-triplet comparison" method in ADASS, wherein triplet of domains is compared with other triplets in a pair-wise comparison of two domain architectures. Different events in the triplet comparison are scored as per a scoring scheme and an average pairwise distance score (Domain Architecture Distance score - DAD Score) is calculated between protein domains architectures. We use domain architectures of a selected domain termed as centric domain and cluster them based on DAD score. The algorithm has high Positive Prediction Value (PPV) with respect to the clustering of the sequences of selected domain architectures. A comparison of domain architecture based dendrograms using ADASS method and an existing method revealed that ADASS can classify proteins depending on the extent of domain architecture level similarity. ADASS is more relevant in cases of proteins with tiny domains having little contribution to the overall sequence similarity but contributing significantly to the overall function.

Key words: Domain architecture, Phylogeny, ADASS, Alignment free domain architecture similarity search.

Background:

Organisms have inherent tendency to innovate and create new proteins and pathways by gene duplication **[1]**, fusion and fission**[2]** through mechanisms like recombination operating at the genomic level[**3**]. This has resulted in a multitude of protein domain architectures having diverse functions within and across species[**4–7**]. Traditionally, proteins are being annotated on the basis of evolutionary relationships, like homology, deduced indirectly from amino acid sequence similarity. Though this strategy works well in the case of single domain proteins, sequence identity would not be sufficient to distinguish between homologues of multi domain proteins. Most of the classifications of multi-domain proteins are based on the sequence similarity between the characteristic functional domains which are common between the proteins[**8**]. However, multi-domain proteins having high sequence similarity in the characteristic domain could still differ functionally, due to the presence of different associated domains. There have been efforts to incorporate the sequence similarity information from such associated domains along with the characteristic domains to deduce the overall protein similarity[**9**]. However due to the differences in length of the associated domains the contribution of these domains to the overall sequence similarity of the proteins vary from domain to domain. The difference in domain architecture of such sequences may not be noticed by simple sequence comparisons. Therefore, detection of homology relationships in multi-domain proteins on the basis of the similarity between the domain architectures is gaining attention.

Although there have been many efforts to compare proteins at the architectural level, they were not very generalized. First of its kind was an alignment-based method **[10],** where an edit-distance method was used to calculate the distance between two domain architectures and was biased for the domain abundance in a protein. Moreover, this method employed dynamic programming algorithm to align domain architectures by considering each domain family as an alphabet forming a "domain sequence". This program provides domain distance between two proteins as number of unmatched domains encountered in such an alignment. However distance between completely unrelated proteins is infinity. Hence such alignment based methods are useful in understanding orthology, but does not perform well in distantly related proteins. Following this, a quantitative measure to compare proteins at the domain architectural level was developed **[11]**. This approach considered three aspects to quantify the similarity between two domain architectures: (1) abundance of shared domains between the two proteins, (2) extent of pair-wise reversal of domain order and (3) amount of duplication of a shared domain. In short, this method utilizes the domain content and order to quantify the similarity between domain architectures. The parameters for this method were optimized for resolving homologues versus non-homologues. Maximum parsimony based methods have also been developed to understand fusion and fission events by processing species trees **[12].** Major follow-up for this approach was a method that considers the complete ancestral tree (not just the species) of each domain and was applied to understand the protein domain architecture evolution **[13].** Most recently, a domain architecture alignment scoring scheme was developed **[14]** based on domain content similarity score that was reported previously **[15].** Essentially, most of these methods examine the generation of alignment based on possible scoring of similarities. Here, we introduce a generalized alignment free method to calculate the distance between two domain architectures. In this paper we present an algorithm that considers domain architecture level similarities between sequences and distinguish proteins from their homologues, classify homologues in to sub-clusters and most importantly detect domain architecture similarity between proteins that are seemingly unrelated at the sequence level.

Methodology:

Algorithm

ADASS algorithm (Alignment free Domain Architecture Similarity Search) compares architectures based on a dataset dependent distance score **(Figure 1).** The algorithm considers individual domains in the domain architecture of a protein as discrete units and computes a distance score. Each of the domain architectures will be divided into triplets and compared with triplets from the other domain architecture. Domain neighbour-hood information is assessed after establishing an exact domain match for central domain of a triplet. Using a relative empirical scoring scheme **(Figure 2),** a score is assigned for each triplet compared. The scores are based on events like complete match, domain duplication, domain shuffle, partial match and no match. The scores for all triplets are summed as Pairwise Distance Score (PDS). The PDS is normalized with the length of both domain architectures (L_1 and L_2) and the maximum triplet score to obtain a Domain Architecture Distance (DAD) score. A DAD score is reported for all domain architecture pairs provided as input. Thus, ADASS scores the domain architecture pairs in such a way that those differing by few domains (architecturally similar) acquire a low DAD score where as those differing by many domains either in number or in order (architecturally dissimilar) acquire a high DAD score .

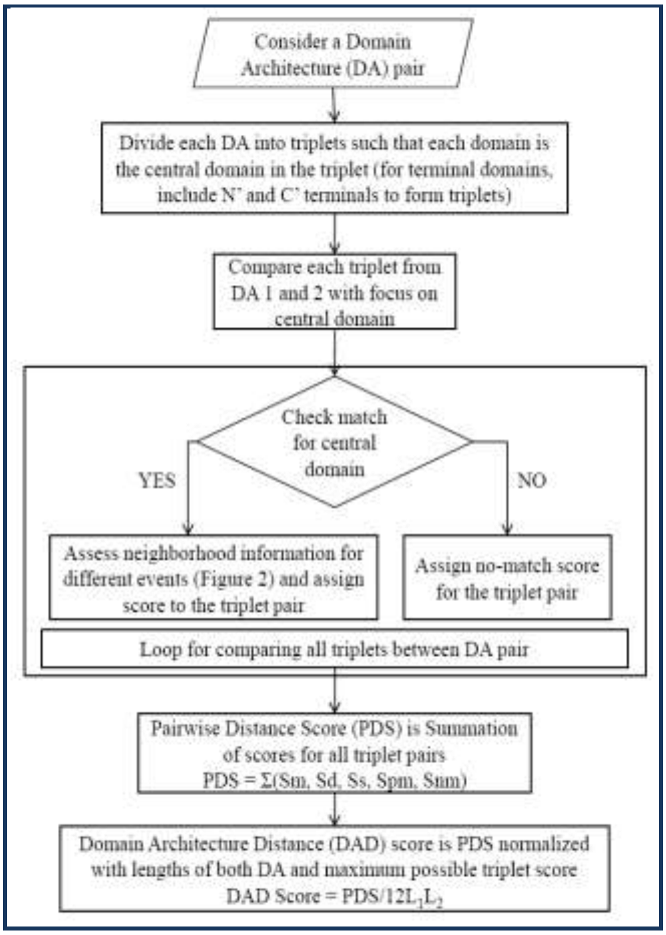
Consider a Domain Architecture (DA) pair
Divide each DA into triplets such that each domain is the central domain in the triplet (for terminal domains, include N' and C' terminals to form triplets)
Compare each triplet from DA 1 and 2 with focus on central domain
Check match for central domain
YES
NO
Assess neighborhood information for different events (Figure 2) and assign score to the triplet pair
Assign no-match score for the triplet pair
Loop for comparing all triplets between DA pair
Pairwise Distance Score (PDS) is Summation of scores for all triplet pairs
PDS = Σ(Sm, Sd, Ss, Spm, Snm)
Domain Architecture Distance (DAD) score is PDS normalized with lengths of both DA and maximum possible triplet score
DAD Score = PDS/12L1L2

Figure 1: All domain architecture (DA) pairs are compared pairwise to obtain Domain Architecture Distance (DAD) score. Triplets are assessed for different events to give scores Sm – match score, Sd – duplication score, Ss – shuffle score, Spm – partial-match score, Snm – no-match score. L1 and L2 are lengths of domain architectures considered in a pairwise comparison.

Domain architecture Datasets

ADASS algorithm was evaluated using five types of domain centric datasets **(Figure 3).** A centric dataset contains domain architectures for a specific domain which is termed as centric domain. These architectures are screened on the basis of presence of domain duplicates or presence in different organisms or the length of architecture. Dataset1, a mixed dataset of heterogeneous (varying length) domain architectures belonging to two important and evolutionarily unrelated protein families - protein kinases (Pkinase - PF00069) and helicases (Helicase_C – PF00271) from human, demonstrates the performance of the algorithm in distinguishing homologous and non- homologous sequences on the basis of domain architectures.

Datasets 2 to 5 were created to estimate the efficiency of the algorithm in handling widely different architectures of a common domain arising from different levels of taxonomy (Datasets2 and 3; for the taxonomy distribution see **Table 1 (see supplementary material)** and from same level of taxonomy (Datasets4and 5; human genome). The domain boundary definitions of these architectures were according to the Pfam database (release 24) **[8].** Dataset 2 consists of 20 different homogeneous (same length) domain-centric datasets (For all datasets, domains corresponding to each centric dataset see **Table 1 (see supplementary material**) devoid of duplications. Dataset 3 has 13 different homogeneous domain centric datasets with duplicates. Dataset4 consists of three different homogeneous paralogous (all three from human genome) domain-centric datasets without duplicates or repeats. Dataset 5 has 11 different homogeneous paralogous domain-centric datasets (all of them are from human genome) and include duplication events.

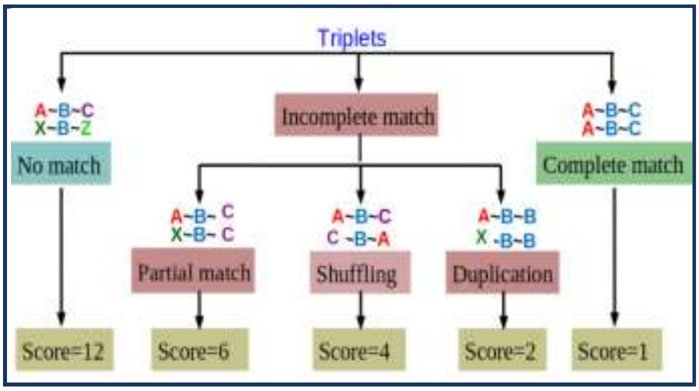

Figure 2: Diagrammatic representation of scoring scheme. Examples of different cases like No match (both the neighbor domains are not matching), incomplete match (one of the neighbor domain matches), and complete match (both the neighbor domains match) are depicted using domains A, B, C, X and Y and their combinations.

Sequence Datasets

To test the effectiveness DAD score in detecting similar domain architectures, pair-wise amino acid sequence identity was considered as a standard. For every centric dataset, the corresponding sequence dataset was generated by considering only single representative sequence for each of the domain architecture present in Pfam (release 24) **[8].** The representative sequences of domain architectures in a domain-centric dataset belong to different organisms ranging from archaebacteria to human. The taxonomic source of sequences is summarized in **Table 1 (see supplementary material).**

Hierarchical Clustering and Tree Construction

Sequence identity benchmark of 40% is widely accepted to classify proteins into homologues or non-homologues. Since such generalized benchmarking of DAD score is inappropriate due to its dependency on the nature of dataset, a hierarchical clustering method was adopted to segregate the domain architectures in a better way on the basis of architectural similarity. The average DAD score (the ADASS computed pair-wise distances) for all possible pairs of architectures were passed through hierarchical clustering algorithm (Neighbor Joining (NJ) method) to generate dendrograms using PHYLIP package **[16].**

Performance Evaluation

Performance of algorithm was assessed by comparing the pair-wise Domain Architecture Distance score (DAD score) with the pair-wise sequence identity obtained using MatGAT tool in various datasets **[17].** DAD score versus sequence identity was plotted as a scatter plot. The mid-range of DAD scores in a plot was considered as the divider for the architecture pairs into low and high DAD score pairs. Sequence identity of 40% was considered as the basis of separation between pairs that are evolutionarily related (>40%) and unrelated (<40%). Thus, the plot was divided in to four quadrants (Quadrant 1 through 4). Quadrants 1(high identity, Low DAD score) Quadrant 3 (low identity, high DAD score) and Quadrant 4 (low Identity, low DAD score) included true positives, whereas Quadrant 2 (high identity, high DAD score) included domain architecture pairs that are considered as false positives (since the high identity pairs are expected to have poor distance score). Pairs falling into Quadrant 4 were found to be cases of proteins that are evolutionarily diverged at sequence level, yet having similar domain architectures. To assess the performance of the algorithm, Positive Predictive Value (PPV) was calculated and was considered as test statistics for the hypothesis:
H0: PPV<DAD ScoreMin; H1: PPV >= DAD ScoreMin
P values for different datasets were calculated using R program.

PPV=Number of True positives/ (Number of True positives+ Number of False positives)

Architecture-based clustering diagrams were generated from the corresponding sequence.

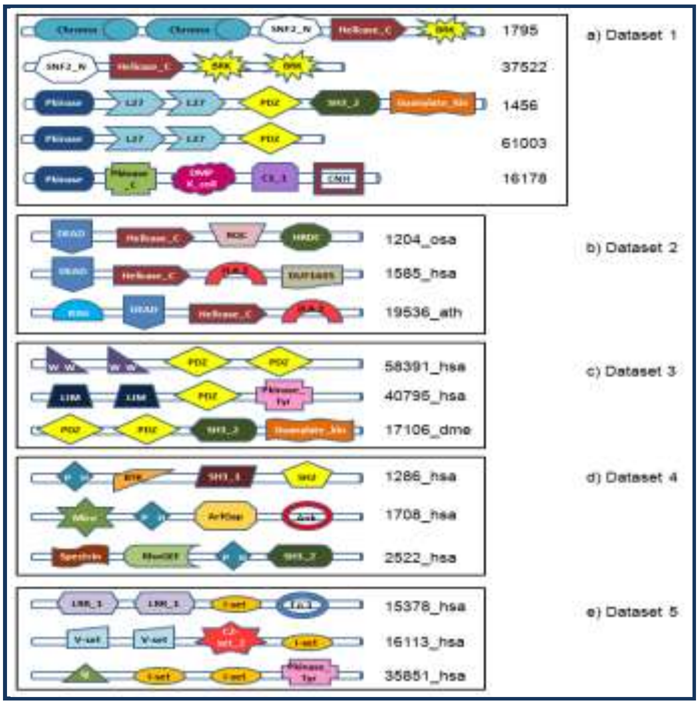

Figure 3: Diagrammatic representation of centric datasets analyzed. Few sample domain architectures from **a)** Pkinase

domain centric and Helicase domain centric architecture datasets that form the dataset5; **b)** DEAD domain centric dataset that belong to dataset1, without duplicates or repeats; **c)** PDZ domain centric dataset that belong to dataset2, i.e. with duplicated domains allowed; **d)** PH domain centric dataset that belongs to dataset3, i.e., without duplicate or repeats; **e)** I-set domain centric dataset that belongs to dataset4, i.e. with duplicated domains allowed.

Results & Discussion:

DAD score distinguishes homologues from non-homologues in more functionally relevant fashion than sequence based approach

Domain architectures, from protein families – Pkinases and Helicases, constituting dataset1 were selected in such a way that the families were mutually exclusive in terms of their domain contents (i.e. no common domain between the families) **(Figure 3a).** All domain architecture pairs formed between members of the same family obtained a DAD score <1.0 indicating the similarities between them where as the pairs formed between families acquired the DAD maximum score of 1.0, pointing to the complete dissimilarity arising from the absence of common domains between families **(Figure 4a).** This highlights the potential of ADASS in distinguishing multi-domain protein homologues from non-homologues having no common domain between them. However the DAD score cut off for categorizing homologues and non-homologues would depend up on the diversity in the domain architectures of the dataset in question.

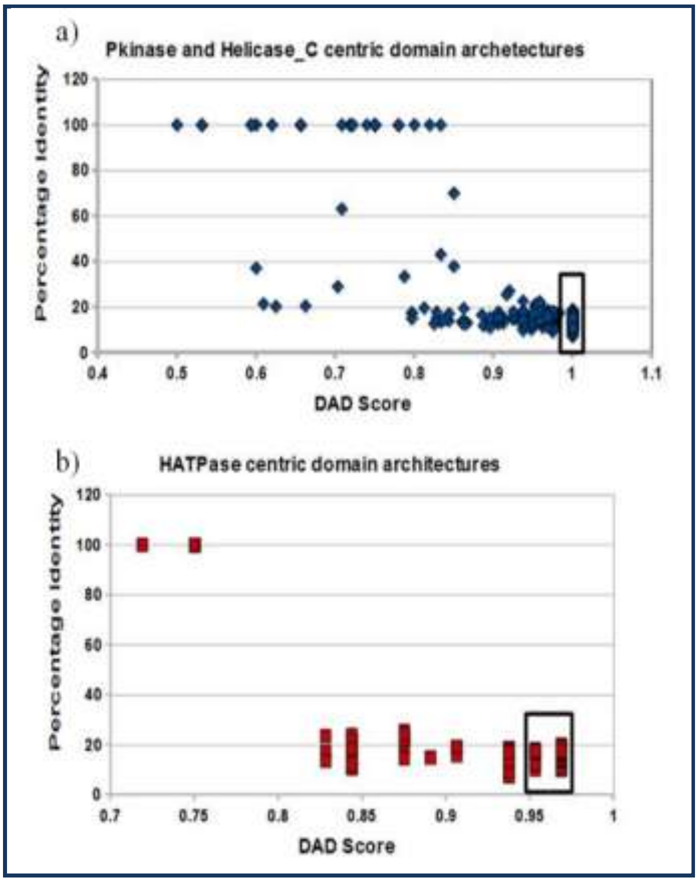

Figure 4: DAD score Vs Percentage identity plot for a) dataset 1. The pairs formed between the completely unrelated Pkinase and Helicase_C obtained the highest DAD score (shown inside box); **b)** HATPase dataset. The pairs formed between architectures having either HisKA or DNAgyraseB along with HATPase obtained highest DAD score (shown inside box)

A hierarchical clustering using DAD score could bring the closest architecture pairs together and distant ones farther in a tree diagram which is more informative to categorize similar architectures and dissimilar ones. DAD score based domain architecture trees, can be used to distinguish homologous sequences from completely unrelated (without any common domain) non-homologous sequences. This is very much evident from the DAD score based clustering diagram **(Figure 5a).** Pkinases and helicases from human, where all protein kinases clustered together and all helicases clustered separately suggesting strongly that the DAD score is sufficient to differentiate the homologues from non-homologues without using sequence information. A full length sequence-based dendrogram **(Figure 5b)** failed to cluster the proteins architectures in to two separate families.

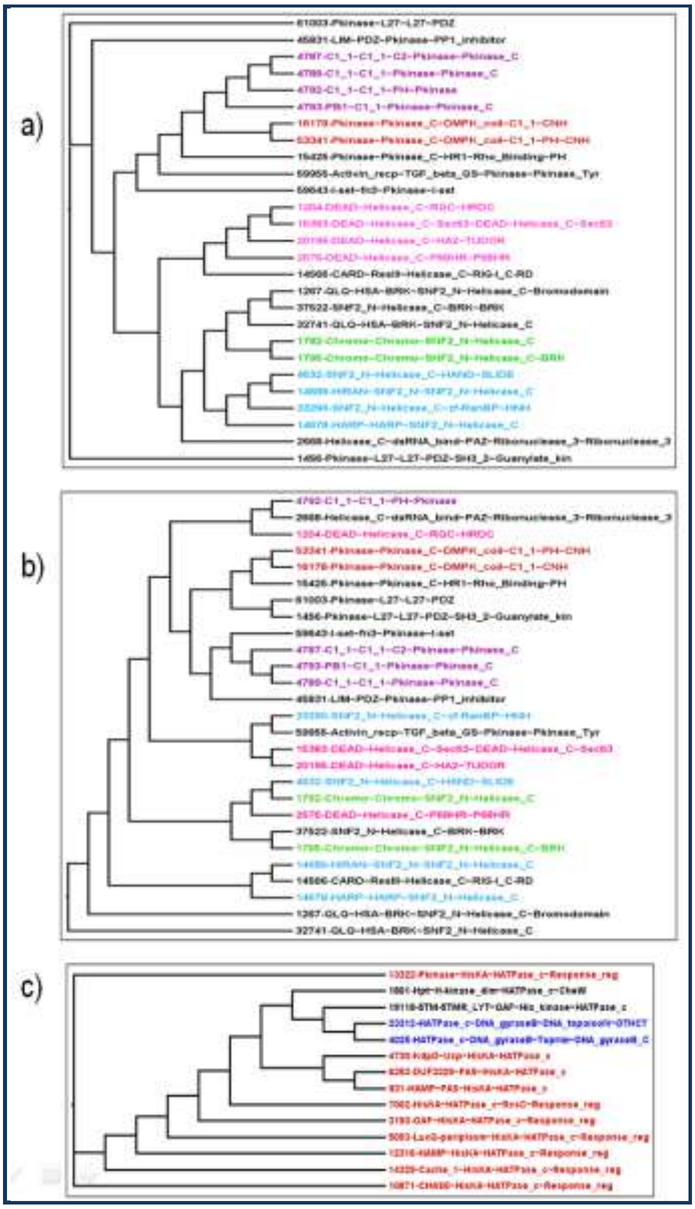
a)
61003-Pkinase-L27-L27-PDZ
45831-LIM-PDZ-Pkinase-PP1_inhibitor
4787-C1_1-C1_1-C2-Pkinase-Pkinase_C
4789-C1_1-C1_1-Pkinase-Pkinase_C
4792-C1_1-C1_1-PH-Pkinase
4793-PB1-C1_1-Pkinase-Pkinase_C
16179-Pkinase-Pkinase_C-DMPK_coil-C1_1-CNH
53341-Pkinase-Pkinase_C-DMPK_coil-C1_1-PH-CNH
15426-Pkinase-Pkinase_C-HR1-Rho_Binding-PH
59955-Activin_recp-TGF_beta_GS-Pkinase-Pkinase_Tyr
59643-I-set-fn3-Pkinase-I-set
1204-DEAD-Helicase_C-RQC-HRDC
16363-DEAD-Helicase_C-Sec63-DEAD-Helicase_C-Sec63
20198-DEAD-Helicase_C-HA2-TUDOR
2575-DEAD-Helicase_C-PSBHR-PSBHR
14506-CARD-ResIII-Helicase_C-RIG-I_C-RD
1267-QLQ-HSA-BRK-SNF2_N-Helicase_C-Bromodomain
37522-SNF2_N-Helicase_C-BRK-BRK
32741-QLQ-HSA-BRK-SNF2_N-Helicase_C
1792-Chromo-Chromo-SNF2_N-Helicase_C
1795-Chromo-Chromo-SNF2_N-Helicase_C-BRH
4632-SNF2_N-Helicase_C-HAND-SLIDE
14689-HIRAN-SNF2_N-SNF2_N-Helicase_C
33294-SNF2_N-Helicase_C-zf-RanBP-HNH
14879-HARP-HARP-SNF2_N-Helicase_C
2688-Helicase_C-dsRNA_bind-PAZ-Ribonuclease_3-Ribonuclease_3
1456-Pkinase-L27-L27-PDZ-SH3_2-Guanylate_kin
b)
4792-C1_1-C1_1-PH-Pkinase
2688-Helicase_C-dsRNA_bind-PAZ-Ribonuclease_3-Ribonuclease_3
1204-DEAD-Helicase_C-RQC-HRDC
53341-Pkinase-Pkinase_C-DMPK_coil-C1_1-PH-CNH
16179-Pkinase-Pkinase_C-DMPK_coil-C1_1-CNH
15426-Pkinase-Pkinase_C-HR1-Rho_Binding-PH
61003-Pkinase-L27-L27-PDZ
1456-Pkinase-L27-L27-PDZ-SH3_2-Guanylate_kin
59643-I-set-fn3-Pkinase-I-set
4787-C1_1-C1_1-C2-Pkinase-Pkinase_C
4793-PB1-C1_1-Pkinase-Pkinase_C
4789-C1_1-C1_1-Pkinase-Pkinase_C
45831-LIM-PDZ-Pkinase-PP1_inhibitor
33290-SNF2_N-Helicase_C-zf-RanBP-HNH
59955-Activin_recp-TGF_beta_GS-Pkinase-Pkinase_Tyr
16363-DEAD-Helicase_C-Sec63-DEAD-Helicase_C-Sec63
20198-DEAD-Helicase_C-HA2-TUDOR
4632-SNF2_N-Helicase_C-HAND-SLIDE
1792-Chromo-Chromo-SNF2_N-Helicase_C
2575-DEAD-Helicase_C-PSBHR-PSBHR
37522-SNF2_N-Helicase_C-BRK-BRK
1795-Chromo-Chromo-SNF2_N-Helicase_C-BRH
14689-HIRAN-SNF2_N-SNF2_N-Helicase_C
14506-CARD-ResIII-Helicase_C-RIG-I_C-RD
14879-HARP-HARP-SNF2_N-Helicase_C
1267-QLQ-HSA-BRK-SNF2_N-Helicase_C-Bromodomain
32741-QLQ-HSA-BRK-SNF2_N-Helicase_C
c)
13022-Pkinase-HisKA-HATPase_c-Response_reg
1861-Hpt-H-kinase_dim-HATPase_c-CheW
19118-STM-STMR_LYT-GAF-His_kinase-HATPase_c
33012-HATPase_c-DNA_gyraseB-DNA_topoisoIV-DTHCT
4026-HATPase_c-DNA_gyraseB-Toprim-DNA_gyraseB_C
4730-KdpD-Usp-HisKA-HATPase_c
6262-DUF2229-PAS-HisKA-HATPase_c
831-HAMP-PAS-HisKA-HATPase_c
7062-HisKA-HATPase_c-RsbC-Response_reg
3193-GAF-HisKA-HATPase_c-Response_reg
5083-LuxQ-periplasm-HisKA-HATPase_c-Response_reg
12316-HAMP-HisKA-HATPase_c-Response_reg
14229-Cache_1-HisKA-HATPase_c-Response_reg
10871-CHASE-HisKA-HATPase_c-Response_reg

Figure 5: a) Domain architecture based dendrogram of mixed dataset5; **b)** Sequence based clustering of mixed dataset 1; **c)** Domain architecture based dendrogram of HATPase proteins.

DAD score can sub divide protein family members into functionally close sub-groups

Multi-domain protein families often possess more than one common domain between them. For instance many architectures containing HATPase domain has additional domains like HisKA and DNAgyraseB. Architecture pairs having either HisKA or DNAgyraseB along with HATPase invariably obtained DAD score equal to or more than 0.95 whereas all the pairs of architectures having both HisKA and DNAgyraseB, the score was <0.95 **(Figure 4b).** The DAD score based hierarchical clustering showed a clear distinction between HisKA domain containing HATPases and other domain architectures that do not contain HisKA domain **(Figure 5c).** This clearly demonstrates ability of ADASS to distinguish not only homologous proteins, but also the ability to categorize a family of domain architectures in to different subfamilies based on the domain content and organization.

Passing low identity sequences of multi-domain proteins through hierarchical clustering might not identify subtle differences in domain architecture that will have great impact on the functional aspects like molecular mechanism or pathway involved. In the case of helicase containing domain architectures, even though the catalytic domain was conserved (DEAD and Helicase_C), due to poor over all sequence identity, these architectures were not clustered together in the sequence dendrogram **(Figure 5b, coloured pink).** However, the DAD score based dendrogram clustered all such architectures together **(Figure 5a, coloured pink).** A similar trend was observed in the case of SNF2 and Helicase_C containing domain architectures as well **(Light blue in Figure 5a and 5b).** This suggests that domain architecture based dendrograms can define subgroups within protein families, which are functionally closer due to the presence of common promiscuous domains. Purely sequence identity based clusters might not capture such functional similarity within the architectures of a family.

Proteins with low sequence identity yet having similar domain architectures are identified

Some architecture pairs, which have poor sequence identities (<40%) obtained low DAD scores **(members of Quadrant 4 in Figure 6a-6e).** These are the most interesting cases since such architecture level similarities cannot be detected by simple sequence similarity measures. Detecting similarity between sequences with low sequence identities is very important from a functional and evolutionary point of view. In the datasets analyzed here, the proportion of such cases was from 0.73% to 2.09% of the total number of pairs analyzed. Even though the frequency of low identity pairs falling in this category is very low **(Figure 6a-6e)**, ADASS could detect such cases with 100% efficiency in all the five datasets analyzed. For instance, the Pfam architectures 4787 (C1_1~C1_1~C2~Pkinase~PkinaseC) and 4789 (C1_1~C1_1~Pkinase~PkinaseC) have poor pair-wise identities of 20.50% and 21.40 % respectively, with the architecture 4792 (C1_1~C1_1~PH~Pkinase). Such poor sequence identities would mislead us to conclude that these proteins are completely unrelated and non-homologous without any common domain between them. To our surprise both these pairs obtained a low DAD score that is well below midrange pointing to the similarity in architectures **(Figure 5a, 7A, 7B)** which was not detectable from the sequence relationship **(Figure 5b).** Gene Ontology (GO) annotations of domains using pfam2go mapping revealed similar molecular function and biological processes in which they are involved **[18].** The three domain architectures differ only by a few domains namely, C2 that is involved in binding to membrane **[19]** and PH that is involved in protein-protein binding **[20]**. The catalytic function Pkinase activity- here is conserved though the substrate specificities and molecular mechanism or localization are different. Thus ADASS can be of use in functional annotation at greater details.

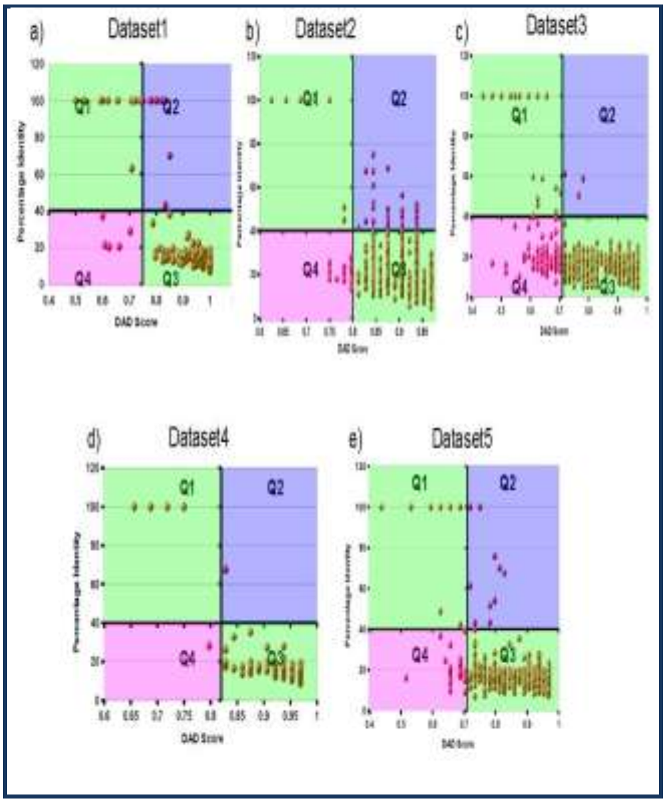

Figure 6: DAD score Vs Percentage identity of individual datasets **(6a-6e).**

In the above mentioned example the component domains are of the comparable size. Nevertheless, the component domains of an architecture can differ in size and hence in the contribution to the overall sequence identity with the sequence of another architecture. Such cases, which are seemingly unrelated at the level of sequence comparisons, may be related in the architecture level. This is being explained in yet another example, in the architecture pair, 1792 (Chromo~Chromo~SNF2_N~Helicase_C) and 1795 (Chromo~Chromo~SNF2_N~Helicase_C~BRK) where Chromo, SNF2 and Helicase_C domains are conserved in the same order, but together these three domains form only 1/5th the size of the whole protein sequence. The only different domain BRK contributes to the sequence level differences to a greater extent leading to a very low sequence identity of 20.3%. But the functional similarity between these sequences is quite high due to the architectural conservation as evident from the DAD score **(see architecture based dendrogram, Figure 5a, Figure 7F)** which simple sequence similarity measures would have failed to identify.

Evaluation of algorithm

Positive Prediction Value (PPV) of different datasets

DAD scores for architecture pairs from different centric datasets are plotted against percentage identities of corresponding representative sequence pair **(Figure 6).** Positive Predictive Value of 1 or very close to 1 were obtained for all the five datasets tested in **Table1 (see supplementary material)** and the P-value of PPV for each dataset was highly significant (<0.001) at a confidence level of 0.05. This clearly demonstrates the ability of the algorithm to handle domain architectures of different lengths **(Figure 6a),** domain architectures from different **(Figure 6a and 6c)** as well as same taxonomic levels **(Figure 6d and 6e**). Figure 6c and 6e indicates that the domain architectures containing duplication events can also be effectively categorized using DAD score. Further, the following observations emphasize on the ability of the program to distinguish between related and unrelated domain architectures.

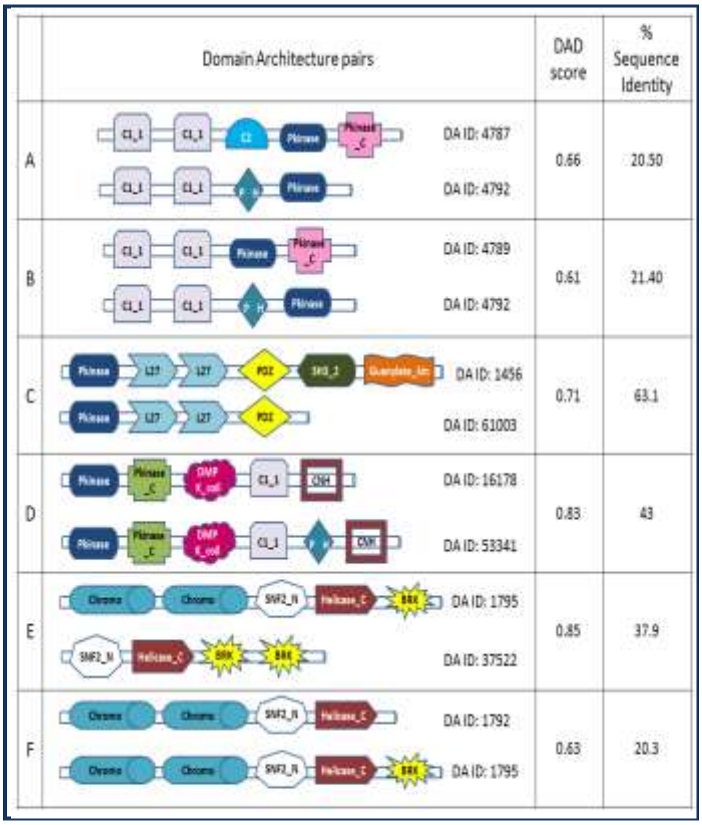

Figure 7: Objective view of domain architecture similarities and differences with respect to DAD score and percentage identity. All architecture pairs (A-F) are from dataset 1. Domain architecture ID (DA ID) mentioned next to each architecture. (Please see text for explanation for each pair).

Evolutionarily related sequences acquired low DAD Score

The sequences with high sequence level similarity are expected to have similar domain architectures. In fact most of the pairs with high sequence level similarity (evident from >40% ID) showed lower DAD score (<mid-range of DAD score) in all the five datasets reflecting the closeness between sequences at the domain architecture level as expected. Such pairs were mostly orthologues belonging to the same family with similar architectures differing only by few domain changes. For example, the domain architectures 1456 (Pkinase~L27~L27~PDZ~Sh3_2~Guanylate_kin) and 61003 (Pkinase~L27~L27~PDZ) of dataset 1 differ only by last two domains, whereas the first four domains are conserved in the same order and hence acquired a low DAD score **(Figure 7C).**

Dissimilar sequences mostly possessed high DAD scores

The datasets analyzed consisted of diverse architectures and hence in general the sequentially dissimilar pairs were the most dominant fraction compared to those with similar pairs **(See graphs in Figure 6a-6e).** Among the low ID sequence pairs, majority (>85%) were segregated to Q3 (Quadrant 3) due to the high DAD score reflecting the architectural differences between the pairs. Thus Quadrant 3, as expected, contained pairs with poor sequence identity that acquired high DAD score pointing to the efficiency of ADASS in identifying dissimilar sequences as dissimilar using an architecture based distance score also.

Comparison with the existing strategies

Previously, there have been attempts to abstract the domain content and order of domain architectures and arrive at domain architecture distance scores and such comparisons can be performed at servers like PDART **[11]** and DAhunter **[21]**, d-omix **[22]** etc. These are useful in comparing homologous architectures, whereas ADASS is a general purpose alignment-free algorithm for estimating similarity relationship between domain architectures irrespective of their homology relationships and is highly sensitive to distinguish proteins with unrelated domain architectures **(Figure 5a).** Unlike other methods developed to detect homology **[11, 14, 15].** ADASS can handle both homologous as well as non-homologous datasets equally well. All the domain architecture distance scores developed so far are based on number of the common domains and order. However ADASS is differing from others as the algorithm systematically scans the neighbourhood of all matching domains.

Tools like CDAC, WDAC, Superfamily etc provide a list of related domain architectures, domain architecture similarity or distance **[23-25].** They do not provide a dendrogram of domain architectures given a set of sequences or domain architectures. PDART server **[11]** is the only publicly available tool that provides dendrograms based on the domain architecture distances. Hence, domain architecture distance based trees developed by ADASS were compared with those from PDART server for a heterogeneous dataset of 16 sample domain architectures of Helicase_C family. In the dendrogram developed using ADASS, all architectures having Chromo domain clustered together (blue box), whereas in PDART dendrogram, the blue box members were clustered along with other architectures with no Chromo domain (orange box members). This shows that the architectures that lack Chromo domain failed to cluster together in the dendrogram obtained using PDART **(Figure 8A and 8B).** In another comparison we found that WDAC server identifies SNF2_N~ResIII~Helicase_C as the best hit (Rank 1) for the query PHD~Chromo~Chromo~ResIII~SNF2_N~Helicase_C **(Figure 8D).** All the hits obtained from WDAC server for this query architecture were used in developing ADASS based dendrogram **(Figure 8C).** According to ADASS the closest architecture to the query is PHD~Chromo~SNF2_N~Helicase_C which according to WDAC acquired only 15th rank.

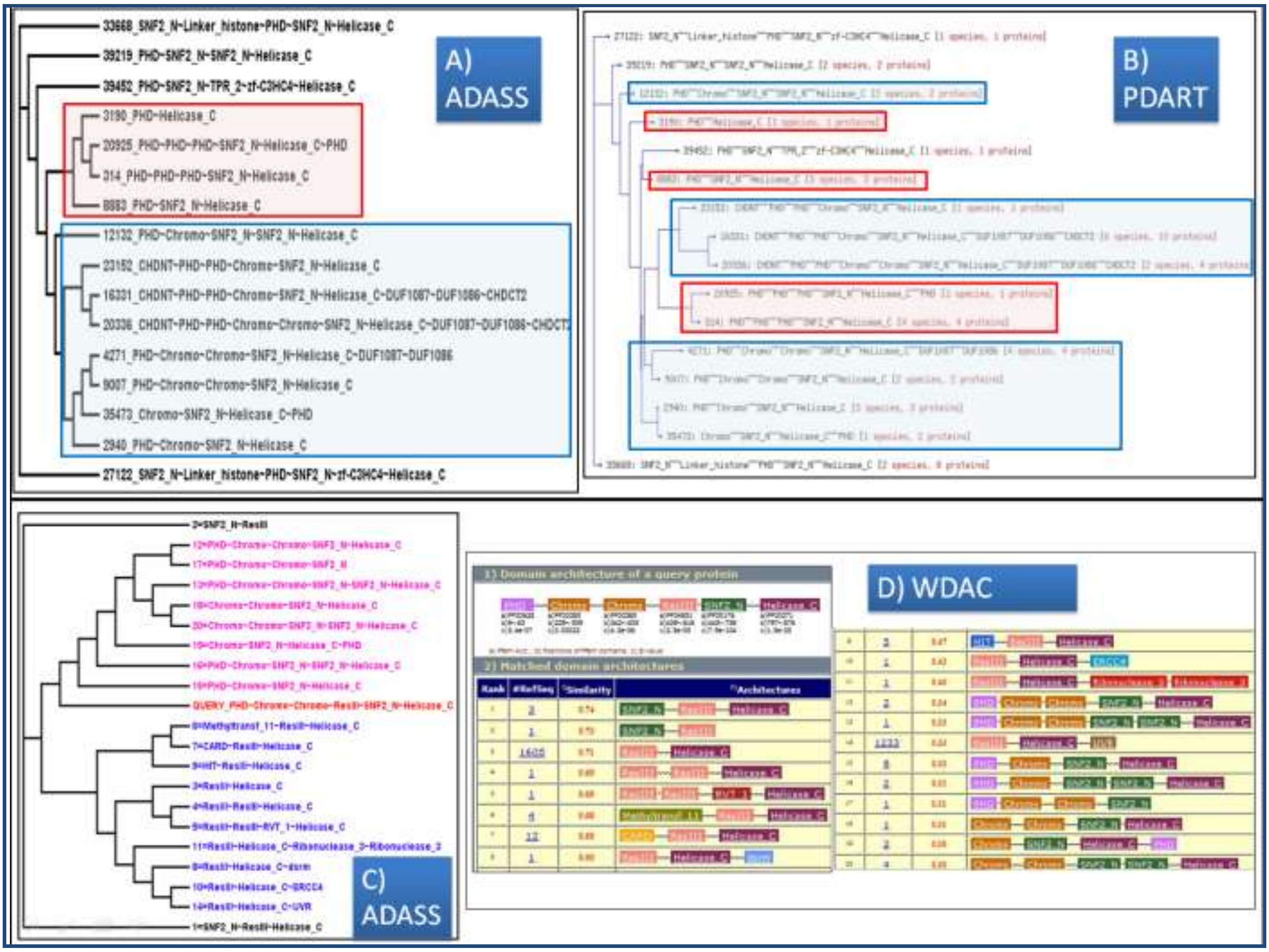
33668_SNF2_N-Linker_histone-PHD-SNF2_N-Helicase_C
39219_PHD-SNF2_N-SNF2_N-Helicase_C
39452_PHD-SNF2_N-TPR_2-zf-C3HC4-Helicase_C
3190_PHD-Helicase_C
20925_PHD-PHD-PHD-SNF2_N-Helicase_C-PHD
314_PHD-PHD-PHD-SNF2_N-Helicase_C
8883_PHD-SNF2_N-Helicase_C
12132_PHD-Chromo-SNF2_N-SNF2_N-Helicase_C
23152_CHDNT-PHD-PHD-Chromo-SNF2_N-Helicase_C
16331_CHDNT-PHD-PHD-Chromo-SNF2_N-Helicase_C-DUF1087-DUF1086-CHDCT2
20336_CHDNT-PHD-PHD-Chromo-Chromo-SNF2_N-Helicase_C-DUF1087-DUF1086-CHDCT2
4271_PHD-Chromo-Chromo-SNF2_N-Helicase_C-DUF1087-DUF1086
9007_PHD-Chromo-Chromo-SNF2_N-Helicase_C
35473_Chromo-SNF2_N-Helicase_C-PHD
2940_PHD-Chromo-SNF2_N-Helicase_C
27122_SNF2_N-Linker_histone-PHD-SNF2_N-zf-C3HC4-Helicase_C
A) ADASS
B) PDART
2=SNF2_N-ResIII
12=PHD-Chromo-Chromo-SNF2_N-Helicase_C
17=PHD-Chromo-Chromo-SNF2_N
13=PHD-Chromo-Chromo-SNF2_N-SNF2_N-Helicase_C
18=Chromo-Chromo-SNF2_N-Helicase_C
20=Chromo-Chromo-SNF2_N-SNF2_N-Helicase_C
19=Chromo-SNF2_N-Helicase_C-PHD
16=PHD-Chromo-SNF2_N-SNF2_N-Helicase_C
15=PHD-Chromo-SNF2_N-Helicase_C
QUERY_PHD-Chromo-Chromo-ResIII-SNF2_N-Helicase_C
6=Methyltransf_11-ResIII-Helicase_C
7=CARD-ResIII-Helicase_C
9=HIT-ResIII-Helicase_C
3=ResIII-Helicase_C
4=ResIII-ResIII-Helicase_C
5=ResIII-ResIII-RVT_1-Helicase_C
11=ResIII-Helicase_C-Ribonuclease_3-Ribonuclease_3
8=ResIII-Helicase_C-dsrm
10=ResIII-Helicase_C-SRCCA
14=ResIII-Helicase_C-UVR
1=SNF2_N-ResIII-Helicase_C
C) ADASS
1) Domain architecture of a query protein
2) Matched domain architectures
Rank
#RefSeq
Similarity
Architectures
D) WDAC

Figure 8: Comparison of domain architecture based dendrogram generated through **A)** ADASS and **B)** PDART. Comparison of domain architecture based dendrogram generated through; **C)** ADASS and **D)** WDAC.

Earlier attempts using maximum parsimony based methods to align homologous proteins have demonstrated an alternate possibility of domain architecture based clustering of proteins**[12]**. However, such methods become difficult to use when the diversity of domain architectures increases in the dataset because one has to device matrices that are huge sized, as big as the number of domains rather than the number of architectures. ADASS does not have such limitation and can categorize architectures of any length and domain diversity.

Conclusion:

Domain architecture level similarities between two proteins, can add value to the sequence similarity-based function annotation and classification in to families or subfamilies. ADASS algorithm compares and classifies protein domain architectures by recognizing similarity between the domain architectures. This is very useful in studying the evolutionary relationship between multi-domain sequences where homology cannot be detected from sequence similarity based approaches alone. ADASS differ from other algorithms in having neighborhood information in its distance score. This approach is novel and has been shown to detect similar architectures and segregate completely dissimilar or partially similar architectures very efficiently enabling subfamily level categorization of domain architectures. An architecture based similarity scoring like ADASS can also provide more insights on the functional similarities or differences compared to simple sequence based similarity measures.

Acknowledgement:
Authors thank NCBS (TIFR), Bangalore for infrastructural support. DPS is supported by NCBS, Bangalore and Sugarcane Breeding Institute, Coimbatore under the Indian Council of Agricultural Research, New Delhi. AJ is supported by Council of Scientific and Industrial Research, New Delhi.

References:

[1] Vogel C *et al. J Mol Biol.* 2005 **345:** 355 [PMID: 15663950]
[2] Kummerfeld SK & Teichmann SA, *Trends in Genetics.* 2005 **21:** 25 [PMID: 15680510]
[3] Buljan M et al. *Genome Biol.* 2010 **11:** R74 [PMID: 20633280]
[4] Schwarz F & Aebi M, *Curr Opn Struc Biol.* 2011 **21:** 576 [PMID: 21978957]
[5] Apic G *et al. J Mol Biol.* 2001 **310:** 311 [PMID: 11428892]
[6] Ekman D *et al. J Mol Biol*. 2005 **348:** 231 [PMID: 15808866]
[7] Levitt M, *PNAS.* 2009 **106:** 11079 [PMID: 19541617]

[8] Finn RD *et al. Nucleic Acids Res.* 2010 **38:** D211 [PMID : 19920124]
[9] Marchler-Bauer A *et al. Nucleic Acids Res.* 2011 **39:** D225 [PMID: 21109532]
[10] Bjorklund AK *et al. J Mol Biol.* 2005 **353:** 911 [PMID: 16198373]
[11] Lin K *et al. Bioinformatics.* 2006 **22:** 2081 [PMID: 16837531]
[12] Fong J *et al. J Mol Biol.* 2007 **366:** 307 [PMID: 17166515]
[13] Forslund K *et al. Mol Biol Evol.* 2008 **25:** 254 [PMID: 18025066]
[14] Forslund K et *al. BMC Bioinformatics.* 2011 **12:** 326 [PMID: 21819573]
[15] Song N *et al. J Comp Biol.* 2007 **14:** 496 [PMID: 17572026]
[16] Felsenstein J, *Cladistics.* 1989 **5:** 164
[17] Campanella JJ et *al. BMC Bioinformatics.* 2003 **4:** 29 [PMID: 12854978]
[18] Hunter S *et al. Nucleic Acids Res.* 2010 **37:** D211 [PMID: 18940856]
[19] Nalefski EA & Falke JJ, *Protein Sci.* 1996 **5:** 2375 [PMID : 8976547]
[20] Burks DJ *et al. J Biol Chem.* 1998 **273:** 31061 [PMID: 9813005]
[21] Lee B & Lee D, *Nucleic Acids Res.* 2008 **36:** W60 [PMID: 18411203]
[22] Wichadakul D *et al.Nucleic Acids Res.* 2009 **37:** W417 [PMID: 19465389]
[23] Geer LY *et al. Genome Res.* 2002 **12:** 1619 [PMID: 12368255]
[24] Lee B & Lee D, *BMC Bioinformatics.* 2009 **10:** S5 [PMID: 19958515]
[25] Gough J & Chothia C, *Nucleic Acids Res.* 2002 **30:** 268 [PMID: 11752312]

Edited by P Kangueane
Citation: **Syamaladevi *et al.*** Bioinformation 9(10): 491-499 (2013)

Supplementary material:

Table 1: Statistical significance of ADASS performance for each dataset. The DAD score range for centric datasets and their mid-range values are used to distinguish True Positives (TP) and False Positives (FP). Positive Prediction Value (PPV) is used to assess the significance of prediction. The sources of domain architectures include mammals (*Homo sapiens,Mus musculus, Ratus ratus, Bos tarus, Canis sp*), Plants and algae (*Arabidopsis thaliana, Zea mays, Oriza sativa, Chlamydomonas sp*),Insects/Invertebrates (*Caenorhabditis elegans, Drosophila melanogaster, Apis mellifera*) Amphibians and fishes (*Frog, Danio rerio*), Birds (*Gallus gallus, Taeniopygia guttata*), Fungi (*Sacchromyces, cerevisae, Shizosacchromyces pombe*), Protozoa (*Plasmodium falciparum, Trypanosoma brucei*), bacteria (*E.coli, Bacillus subtilis, Vibrio fischeri, Mycoplasma pneumonia*) and Archaea (*Thermotoga maritime, Methanocaldococcus jannaschii, Pyrococcus furiosus*). Dataset 1, 4 and 5 contain domain architectures and sequences only from *Homo sapiens*, whereas Dataset 2 and 3 comprises of all above listed organisms.

Dataset	**No. of centric-domain datasets**	**Names of centric domains**	**DA D Score Range, Mid range ***	**PPV***
Dataset1	1	Pkinase & Helicase_C	0.50-1, 0.75	0.98
Dataset2	20	ABC_tran, Ank, Chromo, DEAD, FERM, fn3, HATPase, HisKA, I-set, LRR, MuDR, MULE, PAS, PDZ, PH, PHD, Pkinase_Tyr, Retrotrans, RhoGEF, RVP_2	0.63-0.97, 0.80	0.99
Dataset3	13	Ank, EGF2, EGF_CA, f5-f8, fn3, Helicase_C, I-set, LRR, PDZ, Pkinase_Tyr, TPR1, V-set, WD40	0.44-0.97, 0.71	1.00
Dataset4	3	Helicase_C, PDZ, PH	0.66-0.97, 0.82	1.00
Dataset5	11	Ank, EGF, fn3, Helicase_C, I-set, PDZ, PH, PHD, Pkinase_Tyr, SH3_1, Sushi	0.44-0.97, 0.71	0.95

* Rounded off to two digits after decimal