Evolutionary Expansion and Specialization of the PDZ Domains

O. Sakarya,[1,2,‡] C. Conaco,[1] Ö. Eğecioğlu,[2] S.A. Solla,[3,4] T.H. Oakley,[5] and K.S. Kosik*[,1,6]

[1]Neuroscience Research Institute, University of California, Santa Barbara

[2]Department of Computer Science, University of California, Santa Barbara

[3]Department of Physiology, Northwestern University

[4]Department of Physics and Astronomy, Northwestern University

[5]Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara

[6]Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara

‡Present address: Genetic Systems R&D, Life Technologies, Foster City, CA 94404.

***Corresponding author:** E-mail: kosik@lifesci.ucsb.edu.

Associate editor: Charles Delwiche

Research article

Abstract

PDZ domains are protein–protein interaction modules widely used to assemble membranous signaling complexes including those found in the neuronal synapse. PDZ-containing genes encoded in metazoan genomes vastly outnumber those in prokaryotes, plants, and fungi. By comparing 40 proteomes to track the evolutionary history of the PDZ domain, we observed that the variety of associations between PDZ and other domains expands greatly along the stem leading to metazoans and choanoflagellates. We asked whether the expansion of PDZ domains was due to random or specific sequence changes. Studying the sequence signatures of 58 PDZ lineages that are common to bilaterian animals, we showed that six common amino acid residues are able to classify 96% of PDZ domains to their correct evolutionary lineage. In PDZ domain–ligand cocrystals, four of these "classifying positions" lie in direct contact with the −1 and −3 residues of the ligand. This suggests coevolution of the more flexible regions of the binding interaction as a central mechanism of specialization inherent within the PDZ domain. To identify these positions, we devised two independent algorithms—a metric termed within-clade entropy (WCE) and an average mutual information (AvgMI) score—that both reached similar results. Extending these tools to the choanoflagellate, *Monosiga brevicollis*, we compared its PDZ domains with their putative metazoan orthologs. Interestingly, the *M. brevicollis* genes lack conservation at the classifying positions suggesting dissociation between domain organization in multidomain proteins and specific changes within the PDZ domain.

Introduction

Protein domains are encoded within genes as compact, spatially distinct structures that can be aligned with sequence-similar homologs and are often positioned in a fixed relationship to other domains. Through domain shuffling, gene duplication, and divergence, the number of genes that harbor the domain can expand and, within the expanded gene set, the organizational relationships among domains can change. A particularly versatile domain, the PDZ domain, consists of 70–80 amino acids with a canonical structure that folds into five to six β-strands and two α-helices (Jemth and Gianni 2007). The domain is named for three proteins—postsynaptic density protein (PSD95), *Drosophila* discs large tumor suppressor (DlgA), and zonula occludens-1 protein (ZO-1)—first discovered to share the characteristic sequence (Songyang et al. 1997).

PDZ domains are found in Archaea, Bacteria, and Eukarya. Two bacterial and four yeast PDZ domains have been reported that include the HtrA serine protease gene that is also present in human (Ponting 1997; Chien et al. 2009). Among Archaea and Bacteria, PDZ domains generally are positioned in tandem to peptidase domains, and in the case of the gram-negative bacterial protein DegS, a member of the HtrA family, the PDZ domain binds the C-terminal consensus sequence YxF, where x is any amino acid (Wilken et al. 2004). In another study, *Escherichia coli* DegS is found to capture the C-terminal Z-motif of misfolded proteins and present their N-terminal segments to the protease domain (Krojer et al. 2008). Less is known about the genes and biology of protistan relatives of Metazoa; however, the genome of the choanoflagellate, *Monosiga brevicollis*, has revealed extensive domain shuffling and duplication of its PDZ domains (King et al. 2008).

Metazoans have greatly diversified the architectures of those genes that contain PDZ domains. Proteins containing PDZ domains anchor transmembrane proteins to the cytoskeleton, hold signaling complexes together, and serve as scaffolds for the synapse as well as many other structures (Ponting et al. 1997; Craven and Bredt 1998; Fanning and Anderson 1999; Sheng and Sala 2001). The membrane-associated guanylate kinases (MAGUKs) are among the most widely studied of the metazoan PDZ domain proteins (te Velthuis, Admiraal, et al. 2007), with a characteristically conserved architecture of one or multiple PDZs, a Src homology 3 (SH3) and a guanylate kinase (GK) domain. Along with the expansion and diversification of the PDZ

domains in metazoans came a concomitant expansion of their binding partners. In addition to the most common motif (Class I: X-S/T-X-L/V-COOH), additional PDZ ligand specificity motifs have been identified (Doyle et al. 1996; Nourry et al. 2003; Zhang et al. 2006), with up to 16 distinct specificity classes conserved from worm to human (Tonikian et al. 2008).

In this study, we have analyzed a large number of bilaterian PDZ domains with the intention of discovering positions in their sequences that can predict the domain setting in which the PDZ domain resides. Extending this analysis to more basal organisms suggested a deep evolutionary pathway toward PDZ specification.

Materials and Methods

Collection of PDZ Domains

Protein peptide sequences of *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Gallus gallus*, *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Tetraodon nigroviridis*, *Takifugu rubripes*, *Ciona intestinalis*, *Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *Caenorhabditis elegans*, and *Escherichia coli* were downloaded from Ensembl FTP site (http://www.ensembl.org/), and of *Nemotostella vectensis*, *Monosiga brevicollis*, *Populus trichocarpa*, *Physcomitrella patens*, *Chlamydomonas reinhardtii*, *Ostreococcus lucimarinus*, *Thalassiosira pseudonana*, *Phaeodactylum tricomutum*, and *Phytophthora ramorum* from JGI web site (http://www.jgi.doe.gov/). Yeast genome sequences (*Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Candida glabrata*, *Saccharomyces castellii*, *Ashbya gossypii*, *Kluyveromyces lactis*, and *Kluyveromyces waltii*) were obtained from BROAD genome database (http://www.broad.mit.edu/). *Apis mellifera* and *Arabidopsis thaliana* sequences were obtained from species' genome Web sites (http://www.hgsc.bcm.tmc.edu/projects/honeybee/ and http://www.arabidopsis.org/). *Amphimedon queenslandica* genomic traces and expressed sequence tags were generated as a collaborative genome project with the Joint Genome Institute and are publicly available (http://www.ncbi.nlm.nih.gov/Traces).

Amphimedon queenslandica traces were searched for putative orthology using a complete list of bilaterian PDZ genes, and candidate traces were assembled into genomic contigs using an in-house assembly pipeline (sequential use of MegaBlast [Zhang et al. 2000] for selection of additional traces and PHRAP or PCAP assemblers [Gordon et al. 1998; Huang et al. 2003] for construction of the contigs). Gene intron/exon structures were identified using GenomeScan (Yeh et al. 2001) and GENSCAN (Burge and Karlin 1997) and additional manual inspection. In *A. queenslandica*, a total of 45 unique genes containing 81 PDZ domains were annotated.

All protein sets were searched using HMMER 2.3.2 (http://hmmer.wustl.edu/) with Pfam-curated GA thresholds to identify and extract PDZ domains (*PF00595*). For architectural analyses, genes containing PDZ domains were extracted and full architectures, that is all associated domains within the same PDZ-containing gene, were identified using all Pfam domains. A total of 4149 PDZ domains were identified in the 40 eukaryotic analyzed species.

Identification of Domain Orthologs and Expansion Tree

Orthology between PDZ domain sequences collected as described above was established using EvolMap (Sakarya et al. 2008) and their standard species tree (supported by Philippe et al. (2009). Minimum ortholog similarity threshold of 25% amino acid identity was used. Expansion tree (Sakarya et al. 2008) is the graphical representation of the gene gain/loss events for a large set of sequences analyzed using Evolmap. For detailed technical information on this method, see the associated reference.

In the case of *M. brevicollis* and Metazoan comparisons, where orthology was difficult to establish due to the long divergence time, candidate orthologous domains with low similarity were further evaluated for shared architecture (domain order synteny) to determine the ancestral domain. Some *M. brevicollis* domains, such as GIPC, Syntenin 1, and Syntenin 2, had 50% or more identity with their Metazoan orthologs, along with shared domain architecture and so were readily defined as orthologs. Others, such as dlg, had a conserved complex domain architecture (PDZ∼PDZ∼PDZ∼SH3∼GuanKin), and most domains of the genes were symmetrical best alignments (PDZ3, SH3, and GuanKin), but some of the domains, for example, PDZ-1 and PDZ-2, were candidate orthologs to more than one domain. In this case, both PDZ-1 and PDZ-2 of dlg have an average 45% identity. Evolmap in this case could not conclude PDZ-1 and PDZ-2 as separate loci at the ancestral genome because *M. brevicollis* dlg PDZ-1 and PDZ-2 are 46% identical, whereas Human dlg PDZ-1 and PDZ-2 are on average 54% identical. Due to the presence of other orthologous domains within the gene, we concluded that a PDZ-1 and PDZ-2 of dlg existed in the ancestral species but had diverged and gained their distinctive metazoan classification signatures later in evolution. The Shank-like gene of *M. brevicollis* harbors a PDZ domain along with an architecture unique to Shank gene (Ank-repeat∼SH3∼PDZ∼SAM), but its PDZ domain was identified as an ambiguous ortholog to multiple PDZ genes with different architectures including the Metazoan Shank. Because the orthologs had a similarity above minimum threshold (38% on average) and Ank-repeat-PDZ association is a unique architecture, we concluded that the ancestor of *M. brevicollis* and animals had a Shank gene with a PDZ domain. Thus, to accept or reject genes with 25–40% sequence identity as candidate orthologs, we required a shared unique domain architecture in which the other shared domains were also orthologous.

Positionwise Entropy

Within-clade entropy (WCE) is a metric we devised to analyze variation at each position within members of

orthologous families. WCE and global entropy (GE) for alignment column x are calculated from the following formulas:

$$\mathrm{GE}(x) = -\sum_{i=1}^{n} p(x_i)\log p(x_i), \quad (1)$$

$$\mathrm{WCE}(x) = \frac{1}{m}\sum_{g=1}^{m}\left(-\sum_{i=1}^{n} p_g(x_i)\log p_g(x_i)\right). \quad (2)$$

The entropy GE of all possible amino acids (i from 1 to $n = 20$) is calculated using its probability distribution $p(x_i)$ for column x. This entropy is high if there are balanced observations of several amino acids, whereas it is low if one amino acid is dominant. The WCE finds the entropy for column x within each clade g using the probability distribution $p_g(x_i)$ for finding amino acid i at column x within group g. These WCEs are then averaged over all clades (g from 1 to m, where m is the number of orthologous clades). A low WCE signals a dominant amino acid present in several groups, regardless of which is the dominant amino acid.

Mutual Information

The mutual information (MI) for each position pair (x, y) and the average MI (AvgMI) for each position x are calculated using the following equations:

$$\mathrm{GE}(x,y) = -\sum_{i=1}^{n} p(x_i)\log p(x_i,y_i), \quad (3)$$

$$\mathrm{MI}(x,y) = \mathrm{GE}(x) + \mathrm{GE}(y) - \mathrm{GE}(x,y), \quad (4)$$

$$\mathrm{AvgMI}(x) = \frac{1}{N-1}\sum_{y=1y\neq x}^{N} \mathrm{MI}(x,y). \quad (5)$$

The joint entropy GE(x, y) is based on the joint probability distribution of amino acids for columns x and y. The mutual information MI(x, y) measures the average information about the amino acid distribution in one of the columns gained from knowing which amino acid is found in the other column (Cover and Thomas 1991) The average of $MI(x, y)$ over all columns y (y from 1 to N, $y \neq x$) for a given position x is the average information AvgMI(x) gained about the amino acid distribution in all other columns by knowing which amino acid is found in column x. The quantity AvgMI(x) provides a tool for the identification of classifying positions.

Comparison of MI Between Natural and Artificial Alignments

The significance of MI scores was evaluated by comparing the actual values to those for a set of 1706 artificial PDZ domain sequences generated randomly from the PDZ HMM using HMM-emit (http://hmmer.wustl.edu/). Domain sequences were aligned using MAFFT. The highest gap-containing positions were removed to leave 76 alignment positions that match the original alignment. MI for each alignment pair was calculated on this pseudorandom alignment. The experiment on random sets was repeated three times, and the averaged MI scores were used for comparison with actual MI scores.

To evaluate the MI between the 58 PDZ clades, the MI for each position pair was calculated from a multiple sequence alignment (MSA) of 58 sequences. Each of these sequences represents a clade. The representative sequence is an exemplar randomly selected from the given clade. The experiment was repeated 10 times, randomly selecting every time a new exemplar of each clade. Results were averaged to obtain the MI for each position pair. Secondly, 58 PDZ exemplars were selected randomly over the set of 1706 real domains, each to represent a presumed clade. This procedure was repeated 10 times as well. Results were averaged to obtain the MI for each position pair. The comparison of these two MI matrices provides an unbiased test of positional MI between the clades.

Crystal Structures

PDZ domain–ligand cocrystals of Erbin, PDB:1MFG (Schultz et al. 1998); INAD, PDB:1IHJ (Kimple et al. 2001); MAGI-1, PDB:2I04 (Zhang et al. 2007); NOS, PDB:1B8Q (Tochio et al. 1999); p55, PDB:2EJY (Kusunoki and Kohno 2007); dlg-2, PDB:2G2L (von Ossowski et al. 2006); dlg-3, PDB:1BE9 (Doyle et al. 1996); Shank, PDB:1Q3O (Im et al. 2003); Syntenin, PDB:1V1T (Grembecka et al. 2006); and Syntrophin, PDB:2pdz (Grembecka et al. 2006) were downloaded from the Research Collaboratory for Structural Bioinformatics Protein Data Bank database and visualized using the software VMD (Humphrey et al. 1996).

Yeast Two-Hybrid Mating Screen

Construction and screening of an *M. brevicollis* cDNA library was performed according to the protocol of the Matchmaker library construction and screening kit (Clontech). To create the cDNA prey library, total RNA was reverse transcribed using a poly(dT) primer (CDSIII). cDNA fragments ranging in size from 300 to 3000 bp were introduced into the pGADT7-Rec vector in AH109 yeast. The bait construct consisting of PDZ(1-3) domains of *M. brevicollis* dlg (amino acids 134–473) was cloned into pGBKT7 and transformed into Y187 yeast. Approximately 8.6×10^7 library clones were screened by mating the PDZ bait strain with the prey library. Interacting clones were identified by growth on selective medium with 15 mM 3-amino-triazole and X-α-galactosidase. Plasmids rescued from positively interacting colonies were transformed into yeast and retested by mating with the PDZ bait strain. Recovered sequences were matched to the assembled *M. brevicollis* genome (JGI) using BlastP or BlastX (www.ncbi.nlm.nih.gov/BLAST/) searches. Putative full-length proteins were predicted using GenomeScan (Yeh et al. 2001). C-terminal sequences were verified by 3′RACE using the FirstChoice RLM-RACE kit (Ambion). Recovered sequences of putative PDZ ligands were cloned and used in GST pulldown assays to further confirm interaction with the *M. brevicollis* dlg PDZ domains.

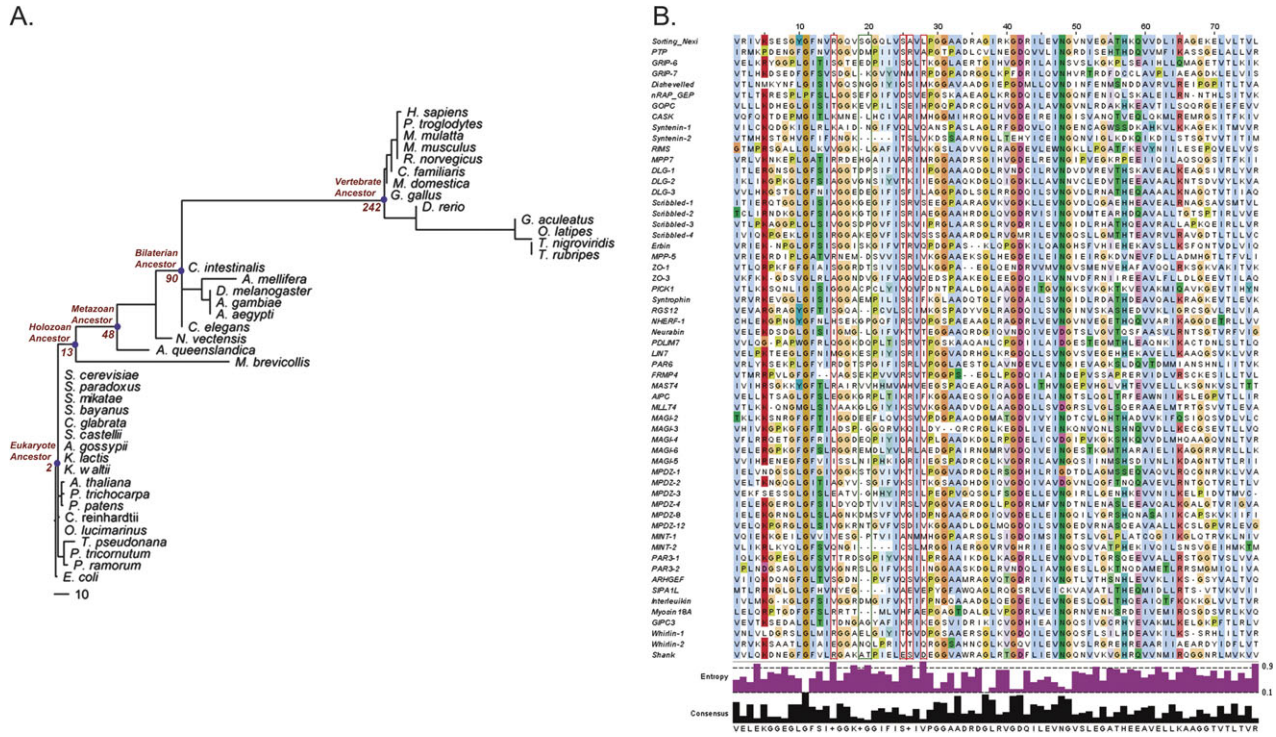

FIG. 1. (*A*) PDZ domain expansion tree. Phylogenetic tree with number of gained PDZ domains shown as branch length. Scale bar corresponds to 10 absolute domain gains (gains minus losses). Gained and lost domains are inferred by symmetrical best alignments and the Dollo parsimony model using the software package Evolmap (Sakarya et al. 2008). (*B*) Alignment of 58 bilaterian PDZ domain families. Each sequence corresponds to the clade consensus of a bilaterian PDZ domain that is strongly conserved in both protostomes and deuterostomes. Globally conserved positions are highlighted according to amino acid type. The bar graphs below the alignment show the global entropy, consensus scores (% identity), respectively, for each position.

Results

Six Amino Acid Positions Define the Bilaterian PDZ Families

To display graphically the natural PDZ domain history, we extracted 4149 PDZ domains from 39 eukaryotic species and the prokaryote, *E. coli*, as an outgroup. From this set, we used domain fasta files and then EvolMap (Sakarya et al. 2008) to find orthologous PDZ domain families. EvolMap utilizes a species tree-based gene clustering method to join all-to-all symmetrical alignments of multiple gene sets in order to infer the gene composition of multiple ancestral genomes and the timings of gene duplications and losses onto evolutionary intervals marked by speciation events. Ninety PDZ domain families were inferred in the bilaterian ancestor (fig. 1*A*), among which 58 were considered well conserved defined as >40% sequence identity among the protostome and deuterostome descendants. These 58 families consisting of PDZ orthologs were referred to as clades. Among the remaining 32 families identified as symmetrical best alignments (see Materials and Methods), the conservation was 25–40%. The 58 bilaterian clades included 1706 modern-day PDZ domains, and the consensus sequence of each clade represents a prebilaterian ancestral estimate for each lineage (fig. 1*B*).

These 1706 domains were aligned using the iterative global multiple alignment option of the MAFFT package (Katoh et al. 2002). In the MSA, positions that contained 30% or more gaps were removed, and 76 positions remained as the global consensus PDZ domain. This MSA shares the same length with the Pfam predicted dlg PDZ-3, which consists of 76 amino acids. Twenty to thirty positions are globally conserved with conservative substitutions, for example, the GLGF loop (fig. 1*B*). In contrast, most of the positions in the alignment are highly variable, for example, position 25 with 14 types of amino acids.

By analyzing variation at each position within members of orthologous families, we devised a metric termed WCE. This value measures the average variation of a position within each clade. For example, although position 25 has high GE, meaning it differs widely among all PDZ domains, it is very well conserved within each clade (low WCE). This was not necessarily true for all variant positions. For example, position 3 had high GE, but also high WCE, suggesting this position varies regardless of the clade to which the PDZ belongs. To find which positions stood out with high GE and low WCE values (i.e., good classifiers of the clades), we generated a scatter plot of average WCE versus GE values for each position over all clades. In this plot, six positions (15, 19, 20, 25, 26, and 28) stood out as outliers of the regression line with confidence interval $\alpha = 0.1$ (fig. 2*A* and Supplementary Table 1, Supplementary Material online). In contrast, if the WCE values were randomly

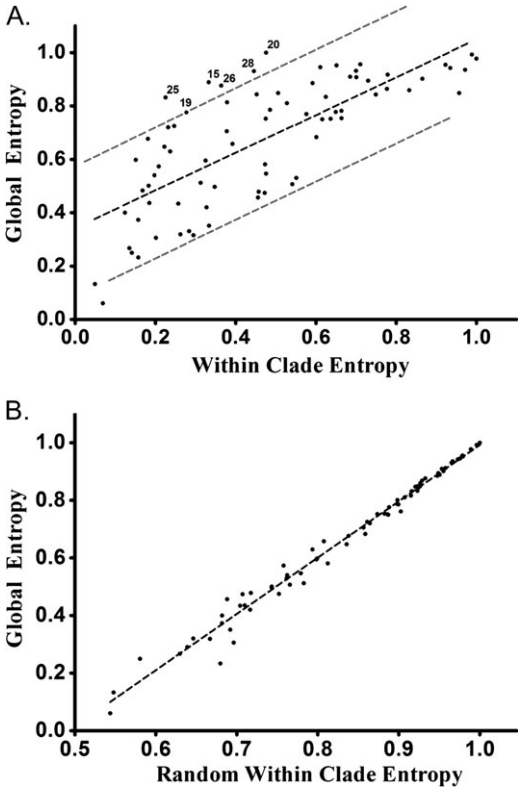

FIG. 2. (*A*) WCE versus GE of the 76 alignment positions. Scatter plot is generated using the average of WCE values (*x* axis) versus the GE values (*y* axis) for each alignment position. Fitted line is found using orthogonal regression and displayed as a dashed black line. Confidence interval ($\alpha = 0.1$) is displayed as dashed lines. Outlier positions (15, 19, 20, 25, 26, and 28) are defined as "classifying." (*B*) Random WCE versus GE of the 76 alignment positions considering the original GE values, WCE values were generated randomly using the restriction 1 > WCE > GE for each of the 58 clades, and averages were calculated for each position.

generated (artificial clades generated from subsets of randomly chosen sequences) with the restriction 1 > WCE > GE, the distribution is observed to be more narrow and without outliers (fig. 2*B*). Thus, the residues that occupy these six positions show extensive variation across all PDZ domains, yet they have been conserved within domain families that predate bilaterians. Gu (1999) has termed this pattern as Type II divergence, wherein an early stage of functional divergence is followed by a long period of purifying selection (Gu 1999; Gu and Vander Velden 2002). This approach implemented in the DIVERGE2 software, unlike WCE, cannot be applied to large numbers of sequences.

To test the predictive merit of the six positions, we devised a procedure that we call "combinatorial classification." In this approach, we calculated the precision with which each combination of positions among the 1706 domains could predict the clade to which a domain belonged. If the selected positions within a specific domain equally matched another clade's consensus signature, then the combination was not considered a precise classifier. For example, human dlg-3 PDZ domain has Valine, Serine, and Phenylalanine at positions 15, 25, and 26, respectively, which only matches the consensus sequence of the dlg-3 clade. Thus, these three positions nonredundantly classify the human dlg-3 to its correct clade. In a three-position combinatorial classification analysis, positions 19, 20, and 25 were found to be the best classifiers with a precision of 81%, whereas the average precision of any three positions (70 300 possible combinations) was 37%. Moreover, any of the 20 three-position combinations of the six positions identified in the WCE versus GE plot were found to have an average clade prediction accuracy of 74%. Combinatorial classification is too slow to find the best four or more position classifiers: testing every four combination of 76 positions would require ~127 billion comparisons (~1 million combinations × 1706 domains × 58 clades), and five combinations would require ~5 trillion. Nevertheless, the six classifying positions from the WCE versus GE plot, when used as a six-position classifier, are able to identify the correct clade with 96% precision. Thus, combinatorial classification corroborated the validity of our classification approach.

Identifying Classifying Positions by MI

Several approaches can quantify the dependence between variables, and among them, information theory provides a more general measure of dependencies than clustering based on Pearson correlation or Euclidean distance (Steuer et al. 2002). When applied to gene expression as a measure of distance, MI groups together genes of known similar function (Butte and Kohane 2000). Variables that are not statistically independent suggest the existence of some functional relation between them. When an approach similar to mutual information was applied to large MSAs, statistically coupled positions were considered a good indicator of thermodynamic coupling (Lockless and Ranganathan 1999). We sought to reevaluate the mutual information within a large MSA after computing the orthologous family organization described above. By definition, mutual information is higher for position pairs that change together and is unaffected by invariant positions. Classifying positions, as defined in the WCE versus GE plot, retain the same amino acid within a clade and change together according to the clade. Therefore, one might expect high MI scores for classifying position pairs.

When the total MI between all pairs, *i*, *j*, in the MSA with 1706 PDZ domains was compared with that of 1706 random PDZ domains generated by HMM-emit (http://hmmer.wustl.edu/), the MI in the original alignment was significantly greater than that of the randomly generated PDZ domains (fig. 3*A*). This suggests a considerable mutual information signal within the natural PDZ alignment. However, the comparison fails to consider background due to the lineage relationships among domains, which could be the basis for the high mutual information (Dunn et al. 2008). Because the clades are much more distantly related (their origin predates the common bilaterian ancestor), we expect the MI to decrease when comparing the consensus sequences of 58 clades. In our case, 58 randomly chosen "exemplar" (Yeates 2005) domains were indistinguishable

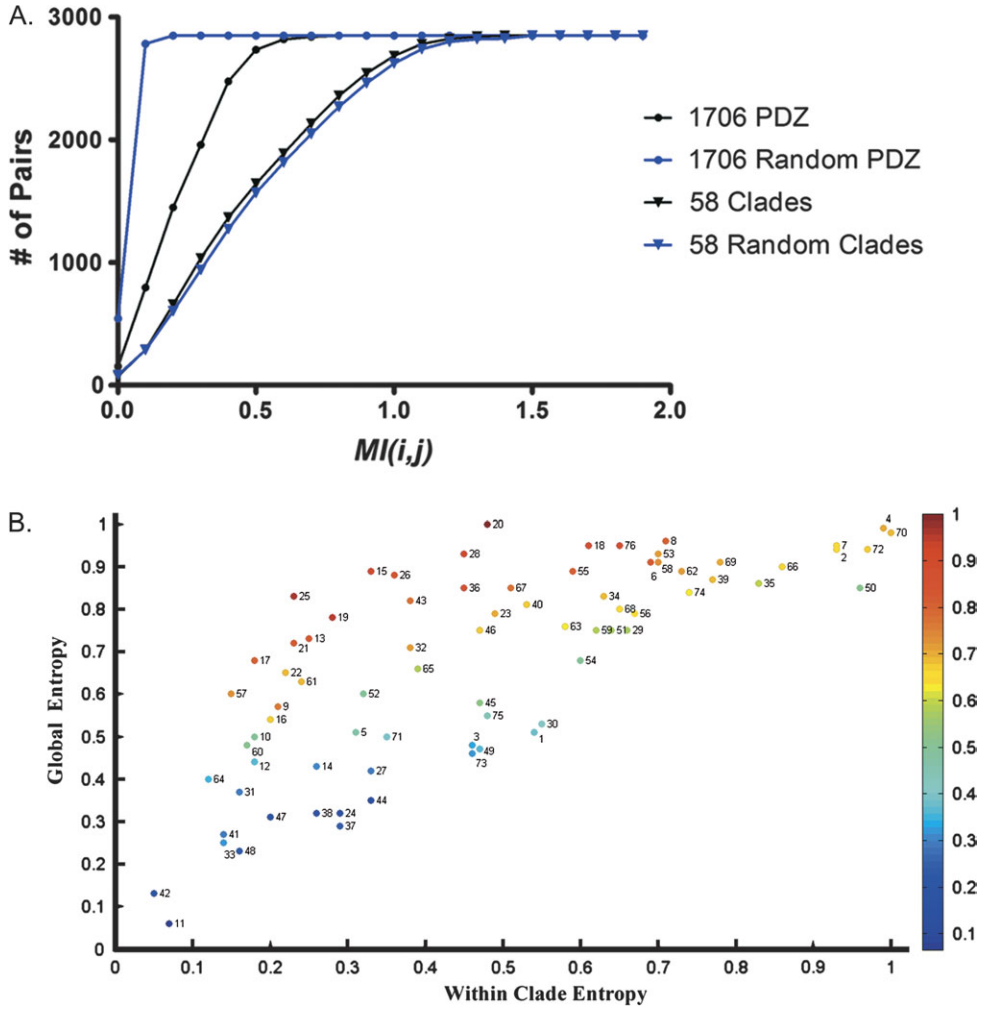

Fig. 3. (*A*) Cumulative MI score distribution in natural and simulated MSAs. "1706 PDZ," and "58 clade" MI values were directly calculated from the original alignments. Three different sets of "1706 random PDZ" domains were generated using HMM-emit and aligned, and their average MI values were calculated. Ten different sets of 58 random exemplar domains were selected from the original 1706 PDZ alignment using a randomized grouping method and their average MI values are calculated (see Materials and Methods for details). (*B*) AvgMI highly correlates with the distance to diagonal of WCE versus GE. AvgMI is displayed as heat map for each position where *x* and *y* axes correspond to WCE and GE, respectively.

in their total MI content from the 58 evolutionarily selected exemplars (fig. 3*A*). If the consensus sequences of the clades were used instead of selecting an exemplar, the total MI between pairs was observed to be even less (data not shown). These results suggest that most of the MI found in the PDZ domain arose from the presence of multiple members of the same lineage within the alignment (see Materials and Methods).

We next calculated the AvgMI of each position *i* to every other position *j* (see Materials and Methods). The correlation of the rank order AvgMI scores, with the rank order WCE versus GE scores (regression line distance) is remarkably high with $r = 0.92$ (fig. 3*B* and Supplementary Table 2, Supplementary Material online). In addition, all six classifying positions ranked very high by AvgMI, and five of them occupy the top five positions. Therefore, average MI among the clades can reveal classifying positions as well as the WCE versus GE plot. In agreement with the exhaustive combinatorial classification results described above, the members of the most predictive triplet—19, 20, and 25—are also the highest ranked AvgMI positions. Therefore, AvgMI method presents an alternative approach to identifying classifying positions, when applied on an alignment of orthologous domains. The methodology we use to identify the domains' classifying positions is summarized in figure 4.

Localization of Classifying Positions in Three-Dimensional PDZ Structures

We assigned the six classifying positions to their specific residues in 10 PDZ ligand cocrystal structures (fig. 5). In almost all crystals, residues in four of the positions (15, 25, 26, and 28) are in direct contact with side chains in the −1 and −3 positions of the PDZ ligand. In Class I PDZ ligands (X-S/T-X-L/V), positions −1 and −3 can be any amino acid and, therefore, point to a role in generating ligand diversity. On the other hand, two classifying positions, 19 and 20, that are among the most predictive of the six classifying positions, are neither in physical contact

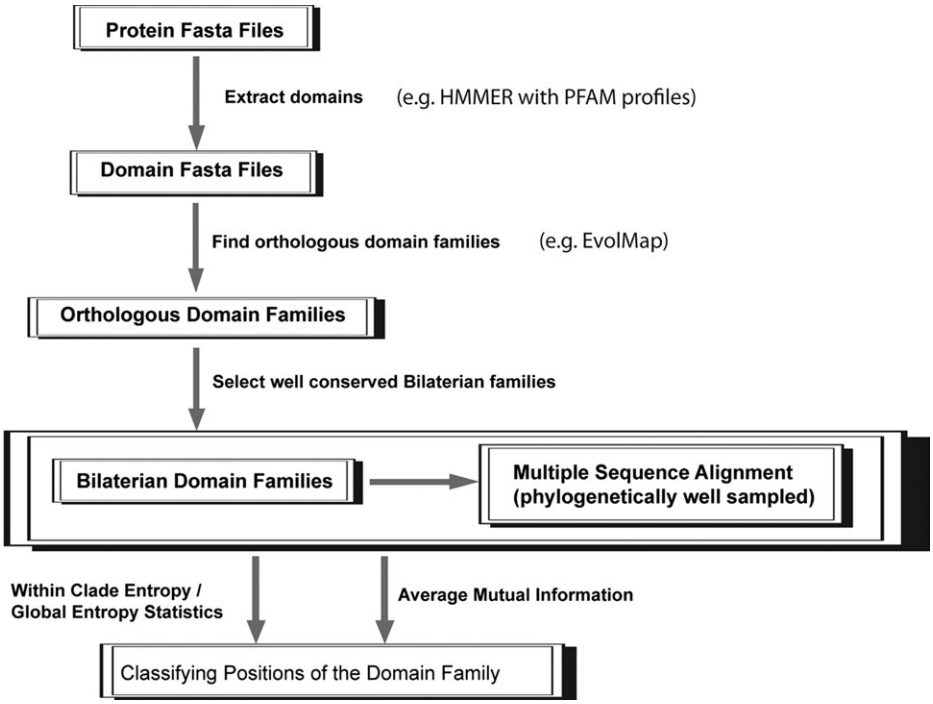

FIG. 4. Summary of the methodology to find classifying domain positions. Orthologous domain families are detected using EvolMap with similarity threshold set as 25% similarity. Bilaterian orthologous families with 40% and above similarity in the two branching clades (deuterostome and protosome) were extracted manually for multiple sequence alignment. Positions of the alignment containing 30% or more gaps were removed. WCE/GE and AvgMI methods are applied on this alignment, with consideration of orthologous groups as described in methods to calculate two statistics for each alignment position in order to find the classifying positions of the domain family.

with the ligand nor do they appear to have a role in dimer formation (data not shown). Curiously, insertions and deletions in the region that include positions 19 and 20 are common.

PDZ Domain Classifying Positions in Early Diverging Metazoa and Pre-Metazoa

The finding that classifying positions have a high AvgMI signal suggests that these positions were selected in the bilaterian ancestor or earlier. We sought the origins of the six classifying positions by analyzing their within-clade conservation among orthologous gene sets in the poriferan, *A. queenslandica*, the cnidarian, *Nematostella vectensis* (Putnam et al. 2007), and the choanoflagellate, *M. brevicollis* (table 1). In *A. queenslandica*, 77% of orthologous PDZ domains have the expected amino acid in at least three of the six classifying positions. In *N. vectensis*, 83% of the orthologous PDZ domains have the expected amino acid in at least three of the six classifying positions. In contrast, *M. brevicollis* has the expected amino acid

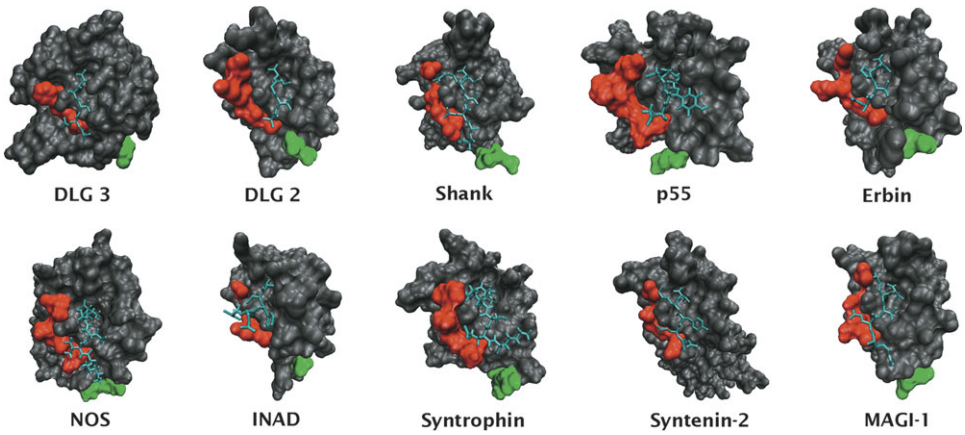

FIG. 5. Positions of classifying residues on PDZ ligand cocrystals. Ten PDZ cocrystals from the PDB database are displayed. The six classifying positions (15, 25, 26, and 28 in red and 19 and 20 in green) are highlighted. Rest of the PDZ domain is highlighted in dark gray and the ligand is highlighted in cyan. Red residues are observed to be almost always in contact with the −1 and −3 positions of the ligand, and occasionally with the −4 position, whereas green residues are not in enough proximity of the ligand for bond formation.

Table 1. Conservation of Classifying Residues in Metazoan Species and *Monosiga brevicollis*.

Species Name	#	p^{15}	p^{19}	p^{20}	p^{25}	p^{26}	p^{28}	Avg	6	5+	4+	3+
Vertebrate	**50**	**0.88**	**0.88**	**0.81**	**0.93**	**0.87**	**0.87**	**0.87**	**0.53**	**0.78**	**0.93**	**0.99**
H. sapiens	44	0.84	0.86	0.84	0.89	0.84	0.77	0.84	0.43	0.73	0.89	1.00
M. mulatta	42	0.81	0.88	0.86	0.90	0.83	0.79	0.85	0.43	0.74	0.90	1.00
R. norwegius	40	0.83	0.88	0.85	0.88	0.83	0.83	0.85	0.45	0.73	0.90	1.00
M. domestica	53	0.91	0.89	0.77	0.96	0.91	0.87	0.88	0.55	0.79	0.96	1.00
G. gallus	56	0.88	0.91	0.80	0.95	0.91	0.91	0.89	0.59	0.80	0.96	1.00
D. rerio	54	0.87	0.91	0.83	0.94	0.91	0.91	0.90	0.63	0.81	0.94	0.98
G. aculeatus	57	0.93	0.89	0.77	0.96	0.88	0.86	0.88	0.54	0.82	0.95	0.98
T. nigroviridis	53	0.92	0.85	0.75	0.92	0.87	0.94	0.88	0.55	0.81	0.94	0.98
T. rubripes	51	0.92	0.88	0.76	0.94	0.86	0.92	0.88	0.61	0.80	0.90	0.98
Invertebrate	**45**	**0.74**	**0.72**	**0.57**	**0.77**	**0.64**	**0.70**	**0.69**	**0.23**	**0.46**	**0.69**	**0.86**
C. intestinalis	31	0.87	0.90	0.68	0.84	0.68	0.52	0.75	0.32	0.61	0.71	0.90
A. mellifera	53	0.77	0.74	0.66	0.75	0.70	0.64	0.71	0.23	0.45	0.74	0.91
D. melanogaster	48	0.79	0.73	0.54	0.77	0.69	0.73	0.71	0.27	0.44	0.69	0.90
A. gambiae	43	0.70	0.77	0.60	0.79	0.63	0.74	0.71	0.26	0.49	0.70	0.88
A. aegypti	42	0.67	0.71	0.62	0.76	0.64	0.71	0.69	0.24	0.48	0.69	0.86
C. elegans	36	0.75	0.64	0.39	0.72	0.56	0.67	0.62	0.11	0.36	0.64	0.78
N. vectensis	46	0.74	0.76	0.61	0.85	0.61	0.72	0.71	0.26	0.52	0.72	0.83
A. queenslandica	30	0.50	0.70	0.50	0.63	0.63	0.37	0.56	0.07	0.27	0.43	0.77
M. brevicollis	13	0.12	0.42	0.35	0.27	0.08	0.19	0.24	0.00	0.00	0.12	0.23

NOTE.—# is the count of clades from the 58 Bilaterian shared families identified to exist within the specified species; $p^{superscript}$ refers to percent identity of the classifying positions within this species averaged over all clades; Avg shows the average conservation of classifying position within this species for all positions; "6" refers to percentage of clades with all six classifying positions conserved; 5+, 4+, and 3+ refer to percentage of clades with 5, 4, and 3 or more classifying positions conserved, respectfully; and vertebrate and invertebrate species refer to the average statistics for a group of species analyzed within this data set.

in only 23% of its six classifying positions among orthologous domains.

Although the metazoan classifying positions were not detectable in *M. brevicollis*, more than half of its PDZ-containing genes share a domain organization with metazoans. Thus, domain architecture, that is, the set of domains in a specified order in a single gene, is dissociated from the classifying positions and may represent a trait acquired along the protometazoan stem. Choanoflagellate PDZ domains expanded vastly as seen by the species tree-based gene clustering method EvolMap (fig. 1A). *Monosiga brevicollis* contains 37 different domain order architectures among 58 PDZ genes in association with 33 other domains. Of these 33 domains, almost none resemble the domain architectures found in bacterial PDZ-containing genes, and more than half are shared with metazoans. Similar to metazoans, PDZ domains in *M. brevicollis* are found in association with SH3, Guan_kin, L27, PH, WW, Ank, SAM, PKinase, C2, and LIM domains. Fourteen PDZ-associated domains in *M. brevicollis* are found within unique architectures that have not been observed in metazoans or other organisms. The MAGUK family is among the metazoan PDZ architectures that are found in *M. brevicollis* but not in other protozoa or plants. A MAGI family protein, which is a subgroup of the MAGUKs, was also reported in the unicellular opisthokont, *Capsaspora owczarzaki* (Ruiz-Trillo et al. 2008).

At least 13 PDZ domains are inferred to have existed at the ancestor of *M. brevicollis* and the metazoans. In the common ancestor of the demosponge *A. queenslandica* and the eumetazoans, at least another 35 PDZ domains became fixed (Table 2). *Amphimedon queenslandica* shares approximately 13 of these with the *M. brevicollis* as distinguishable orthologs. Before the bilaterian ancestor, another 42 PDZ domains became fixed, including several new relationships to tandem domains. Eighteen of these are not found in *N. vectensis*, suggesting that Eumetazoan ancestor had at least 72 PDZ domains. From the bilaterian to the vertebrate ancestor, the number of PDZ domains increased from 90 to 242. Most of this expansion can be attributed to retention of PDZ paralogs that emerged after prevertebrate genome duplications (Dehal and Boore 2005). One of the pair of domains that contributed to the vertebrate expansion is the PDZ/LIM family (te Velthuis, Isogai et al. 2007). In contrast to this expansion within metazoa, the genomes of yeast contain one or two PDZ domains (although more may be discovered in early branching fungi), and the unicellular eukaryotes *Dictyostelium* (Eichinger et al. 2005) and *Tetrahymena* (Eisen et al. 2006) do not contain any PDZ domains. The ancestor of land plants is inferred to have had five PDZ domains, including three associations with specific peptidase domains.

PDZ Domain Ligands in *M. brevicollis*

In bilaterians, *dlg* is a gene family that encodes scaffolding proteins, such as PSD95, that lie at the core of the postsynaptic junction. We screened a *M. brevicollis* poly(dT)-primed cDNA library for peptides that interact with PDZ(1–3) of the *M. brevicollis dlg* gene using the Gal4 yeast two-hybrid (Y2H) system. The screen yielded PDZ ligand consensus sequences at the C-termini of the captured peptides (fig. 6). Interestingly, most of the sequences retrieved belong to the Class I ligand motif (X-S/T-X-L/V), a sequence that departs from the C-terminal consensus sequences observed for ligands of prokaryotic PDZ peptidases. Of the putative ligands that were mapped to *M. brevicollis* genes with metazoan homologs, none bear

Table 2. PDZ Genes of Eukaryote, Holozoan, and Metazoan Ancestors.

Description	Most Common Domain Architecture[a]	Euk	Hol	Met	*Amphimedon Queenslandica, Monosiga brevicollis*[b]
PEPTIDASE S41 SUPERFAMILY	PDZ~Peptidase_S41	1	0	0	NA, NA
HTRA SERINE PROTEASE	TRYPSIN~PDZ	1	1	1	Aq16, Monbr1-7+
GIPC	PDZ	0	1	1	Aq10, Monbr1|39217
SYNTENIN SYNDECAN BINDING	PDZ~PDZ	0	1	1	Aq2, Monbr1|17135
SORTING NEXIN 27	PDZ~PX~RA	0	1	1	Aq34, Monbr1|35789
DISCS LARGE PSD	PDZ~PDZ~PDZ~SH3~GuanKin	0	1	1	Aq31, Monbr1|209
SHANK	Ank-repeat~SH3~PDZ~SAM	0	1	1	Aq3, Monbr1|28170
PRE-MAGUK SUPERFAMILY	PDZ~PDZ	0	1	1	Aq-10+, Monbr1-10+
SAM-PDZ-PH FAMILY	SAM~SAM~PDZ~PDZ~PH	0	1	1	NA, Monbr1|24312
PDZ-LIM FAMILY	PDZ-LIM-LIM	0	1	1	NA, Monbr1|38749
TYROSINE PHOSPHOTASE	Band-41~FA~PDZ~Y-phosphatase	0	0	1	Aq29
DISHEVELLED	DIX~Dishevelled~PDZ~DEP	0	0	1	Aq20
PALMITOYLATED 2 (MAGUK)	L27~PDZ~SH3~Guanylate-kin	0	0	1	Aq5
SCRIBBLE	LRR-repeat~PDZ~PDZ~PDZ~PDZ	0	0	1	Aq11
LIN-7	L27~PDZ	0	0	1	Aq27
PALMITOYLATED 7 (MAGUK)	L27~PDZ~SH3~GuanKin	0	0	1	Aq6
PAR-6	PB1~PDZ	0	0	1	Aq12
PICK1	PDZ~Arfaptin	0	0	1	Aq32
NHERF	PDZ~PDZ	0	0	1	Aq18,19
MAGI	PDZ~WW~WW~PDZ(×5)	0	0	1	Aq9
X11 MINT	PID~PDZ~PDZ	0	0	1	Aq25
CG3402 PDZ PROTEIN	PDZ	0	0	1	Aq13,14,22,42
MAST	DUF1908~Pkin~Pkin-C~PDZ	0	0	1	Aq43
GRIP	PDZ(×6)	0	0	1	Aq35
NOS	PDZ~NO-synthesase~FAD~NAD	0	0	1	Aq24
AFADIN	RA~RA~FHA~DIL~PDZ	0	0	1	Aq28
MYOSIN CONTAINING	PDZ~Myosin-head~IQ	0	0	1	Aq42
WHIRLIN	PDZ-PDZ-PDZ	0	0	1	Aq4

NOTE.—Euk, Eukaryote; Hol, Holozoan; Met, Metazoan; and NA, not applicable

[a] Some genes may have a different architecture than most common architecture shown.

[b] *Amphimedon queenslandica* sequences (AqX) are provided as supplemental data. *Monosiga brevicollis* gene IDs are from JGI v1–filtered gene sets. Large families with multiple members are shown with x+ annotation.

any resemblance to postsynaptic proteins known to interact with PSD95 (table 3).

Although the yeast two-hybrid screen demonstrates the competence of *M. brevicollis* dlg tri-PDZ domain in capturing ligands with the metazoan consensus sequence, it is important to note several limitations of this method. First, the interaction between *M. brevicollis* dlg and the putative ligands may not necessarily be representative of

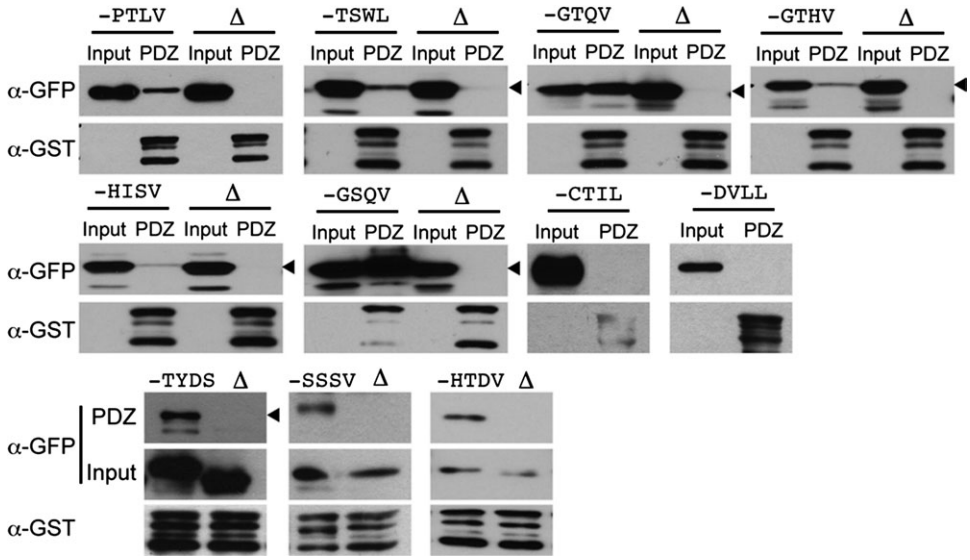

FIG. 6. *Monosiga brevicollis* dlg binds to Class I C-terminal ligands. Putative *M. brevicollis* dlg ligand sequences recovered from a yeast two-hybrid screen are able to interact with the PDZ(1-3) domains of *M. brevicollis* dlg (PDZ) in a GST pulldown assay. Deletion of C-terminal amino acids (Δ) on the ligand abolishes the interaction. Ligand sequences were tagged with green fluorescent protein (GFP) and the PDZ domains were tagged with GST. Input represents 1% of total protein used in the pulldowns. Western blots were probed with antibodies against GFP (α-GFP) or GST (α-GST). Arrowheads indicate the band of interest for tagged ligands.

Table 3. Summary of *Monosiga brevicollis* dlg Yeast Two-Hybrid Screen.

Putative Ligand C-Terminus	Predicted Human Gene Homolog (BlastP/BlastX)	C-Terminus Verified by 3′RACE	Remating Assay	GST Pulldown
-PTLV	APG10-like (NP_113670)	Yes	+	+
-TYDS	Cathepsin S preproprotein (NP_004070)	Yes	+	+
-SSSV	ATP/GTP binding protein-like 4 (NP_116174)	Yes	+	+
-HISV	Transformer 2 beta homolog (NP_004584)	No	+	+
-HTDV	Unknown	No	+	+
-GSQV	Unknown	No	+	+
-GTHV	Unknown	No	+	+
-PTAV	Unknown	No	+	+
-GTQV	Unknown	No	+	+
-TSWL	Unknown	No	+	+
-CTIL	Calcium binding protein 1 (NP_001028849)	Yes	+	−
-DVLL	Unknown	No	−	−

in vivo interaction and will need to be validated for the endogenous proteins. Second, PDZ domains bind to many transmembrane proteins and these interactions may not be readily detected using the Gal4 system that requires transactivation of reporter genes in the nucleus. However, PSD95 also interacts with many nonmembrane proteins, and thus, the Gal4 yeast two-hybrid system was preferable for our purposes. Moreover, cDNA fragments containing cytoplasmic C-termini of membrane proteins will still be detected using this method. For example, the interaction between the C-terminus of plexinB1 and the PDZ domain of RhoGEF was detected using a similar system (Driessens et al. 2002).

Among the genes that encode dlg PDZ ligands in metazoans and also present in *M. brevicollis* are CRIPT, Shaker-like K^+ channel, and PMCA. Although these genes are clearly recognizable orthologs of their metazoan counterparts, they all lack a C-terminal PDZ ligand sequence in the *M. brevicollis* genes. The absence of a PDZ ligand sequence in these orthologous genes suggests that they may not bind PDZ domains. However, this was not directly tested, and they may have some other means of binding to form an orthologous complex. None of these findings speak to whether the ligand sequences in these genes were gained in metazoans or lost in choanoflagellates.

Discussion

One way to classify PDZ domains is by the company they keep. PDZ domains are found in tandem with other domains, including other PDZ domains. In contrast to Bacteria, Archaea, and many single-celled eukaryotes and plants, PDZ domain associations are vastly diverse among animals (see supplementary data, Supplementary Material online). Their paucity in fungi suggests an expansion along the holozoan stem. Although we assumed that independent co-occurence of domain architectures is rare, convergent assembly of similar domains is possible. In all animals, including the demosponge *A. queenslandica*, we found that the particular amino acid in six specific positions within the PDZ domain can predict the gene from which that PDZ domain is derived. In the 600 million years of animal evolution, these positions have remained constant, suggesting that they are under strong purifying selection at a time when large changes in PDZ ligand selection were taking place at the origin of animals. Four of the classifying positions appear to have a role in determining ligand specificity based on their proximity to variable residues in the PDZ-binding sequence. The pervasiveness of these classifying positions throughout the animal kingdom is supported by the observation that human and worm PDZ orthologs have nearly identical ligand specificities (Tonikian et al. 2008). Mutations in the corresponding four residues of the Erbin PDZ domain (β2-4, β3-4, β3-5, and β3-α1-1) were previously shown to change the ligand specificities for -1 and -3 positions (Tonikian et al. 2008). A detailed ligand-binding analysis of PDZ2 and PDZ3 of PSD95 and SAP-97 showed that ligand position -1 is a discriminator of binding between individual PDZ domains (Kurakin et al. 2007).

The significance of the remaining two classifying positions—19 and 20—in the PDZ domain is not obvious. Their very high MI values suggest a deep and highly conserved evolutionary relationship. However, their distance from the ligand-binding pocket makes a direct effect on ligand binding less likely. Using residual high MI after removal of background MI from random noise and phylogenetic effects, Dunn, Wahl, and Gloor (2008) found that the vast majority of their high MI position pairs were in contact and considered these position pairs as coevolving. One possible explanation for the inclusion of positions 19 and 20 in the set of positions with high MI is a mode of structural evolution called conformational epistasis (Ortlund et al. 2007). Because we generally assume that evolutionary pathways pass through functional intermediates (Smith 1970), paired positions may bear conformationally epistatic relationships to each other. As noted by Ortlund et al. (2007), a mutation may create a permissive sequence environment for additional substitutions capable of remodeling the protein over evolutionary time periods and evolving novel molecular interactions.

In a common ancestor of the choanoflagellates/metazoans, many PDZ domains assumed positions in genes that remained conserved within both the choanoflagellate and metazoan lineages. For example, a gene with the domain

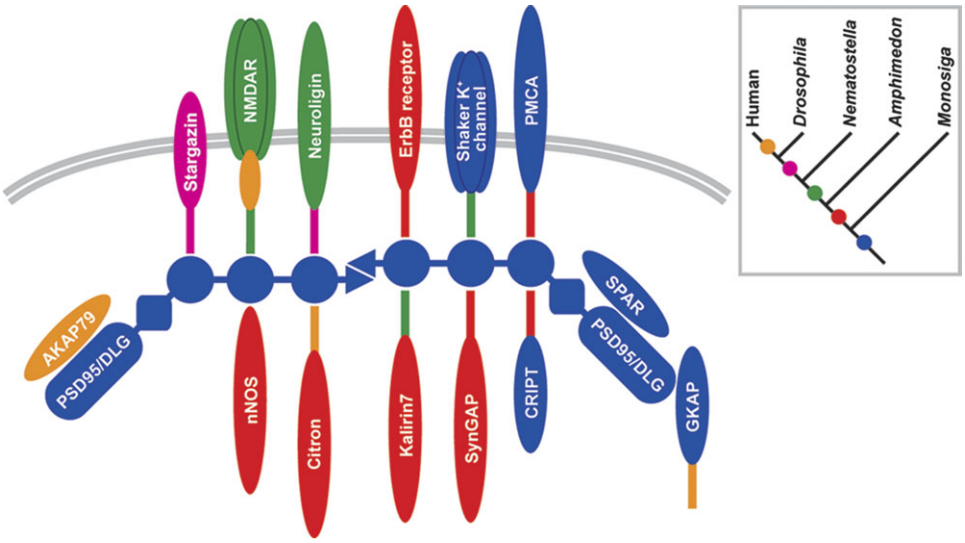

Fig. 7. Origination periods of genes that interact with the dlg postsynaptic scaffold. Postsynaptic genes and their PDZ ligand sequences originated at different times during animal evolution. Colors represent the time of origination as indicated by the species tree on the right. PDZ ligand sequences are represented by short rectangles.

structure of PSD95, that is, three PDZ domains, an SH3 domain, and a GK domain, appeared in a common ancestor of the choanoflagellates/metazoans. Interestingly, this PSD95 ortholog shares the bilaterian PSD95 property of binding to Class I ligands (X-S/T-X-L/V), but lacks conservation of the classifying amino acids at the six classifying positions. This observation suggests a dissociation between the domain structure of PSD95 and the defining sequence features of its PDZ domains along the stems leading to choanoflagellates and metazoa.

Previously, we conjectured that the demosponge, *A. queenslandica*, uses a set of scaffold proteins containing PDZ domains to assemble a structure that resembles the postsynaptic junction (Sakarya et al. 2007). Many of the PDZ scaffolds found in the metazoan ancestor are important components of the modern vertebrate synapse (table 2). A comparison of the postsynaptic genes present in *M. brevicollis*, *A. queenslandica*, and the Bilateria suggests a possible evolutionary assembly path to the postsynaptic complex (fig. 7). The presence of a *dlg* ortholog in *M. brevicollis* and *A. queenslandica* suggests a starting point to search for structures that bear deep homology with synapses.

Supplementary Material

Supplementary Tables 1 and 2 and data are available at Molecular Biology and Evolution online (http://mbe.oxfordjournals.org/).

Acknowledgments

We thank Boris Shraiman, Richard Neher, Pierre Neveu, Bernard Degnan, Kathryn Armstrong, and Bruce Tidor for insightful discussions. Thanks to Harvey Karp for his support of this work.

References

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 268:78–94.

Butte AJ, Kohane IS. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*. 5:418–429.

Chien J, Ota T, Aletti G, Shridhar R, Boccellino M, Quagliuolo L, Baldi A, Shridhar V. 2009. Serine protease HtrA1 associates with microtubules and inhibits cell migration. *Mol Cell Biol*. 29:4177–4187.

Cover TM, Thomas JA. 1991. Elements of information theory. New York: Wiley-Interscience. Chapter 2, p. 13–54.

Craven SE, Bredt DS. 1998. PDZ proteins organize synaptic signaling pathways. *Cell* 93:495–498.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. 3:e314.

Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. 1996. Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* 85:1067–1076.

Driessens MH, Olivo C, Nagata K, Inagaki M, Collard JG. 2002. B plexins activate Rho through PDZ-RhoGEF. *FEBS Lett*. 529:168–172.

Dunn SD, Wahl LM, Gloor GB. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24:333–340.

Eichinger L, Pachebat JA, Glockner G, et al. (97 co-authors). 2005. The genome of the social amoeba Dictyostelium discoideum. *Nature* 435:43–57.

Eisen JA, Coyne RS, Wu M, et al. (53 co-authors). 2006. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biol*. 4:e286.

Fanning AS, Anderson JM. 1999. PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. *J Clin Invest*. 103:767–772.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res*. 8:195–202.

Grembecka J, Cierpicki T, Devedjiev Y, Derewenda U, Kang BS, Bushweller JH, Derewenda ZS. 2006. The binding of the PDZ

tandem of syntenin to target proteins. *Biochemistry* 45: 3674–3683.
Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol*. 16:1664–1674.
Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* 18:500–501.
Huang X, Wang J, Aluru S, Yang SP, Hillier L. 2003. PCAP: a whole-genome assembly program. *Genome Res*. 13:2164–2170.
Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph*. 14:33–38 27–38.
Im YJ, Lee JH, Park SH, Park SJ, Rho SH, Kang GB, Kim E, Eom SH. 2003. Crystal structure of the Shank PDZ-ligand complex reveals a class I PDZ interaction and a novel PDZ-PDZ dimerization. *J Biol Chem*. 278:48099–48104.
Jemth P, Gianni S. 2007. PDZ domains: folding and binding. *Biochemistry* 46:8701–8708.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30:3059–3066.
Kimple ME, Siderovski DP, Sondek J. 2001. Functional relevance of the disulfide-linked complex of the N-terminal PDZ domain of InaD with NorpA. *EMBO J*. 20:4414–4422.
King N, Westbrook MJ, Young SL, et al. (36 co-authors). 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* 451:783–788.
Krojer T, Pangerl K, Kurt J, Sawa J, Stingl C, Mechtler K, Huber R, Ehrmann M, Clausen T. 2008. Interplay of PDZ and protease domain of DegP ensures efficient elimination of misfolded proteins. *Proc Natl Acad Sci USA*. 105:7702–7707.
Kurakin A, Swistowski A, Wu SC, Bredesen DE. 2007. The PDZ domain as a complex adaptive system. *PLoS ONE*. 2:e953.
Kusunoki H, Kohno T. 2007. Structural insight into the interaction between the p55 PDZ domain and glycophorin C. *Biochem Biophys Res Commun*. 359:972–978.
Lockless SW, Ranganathan R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
Nourry C, Grant SG, Borg JP. 2003. PDZ domain proteins: plug and play!. *Sci STKE*. 2003:RE7.
Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317:1544–1548.
Philippe H, Derelle R, Lopez P, et al. (20 co-authors). 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol*. 19:706–712.
Ponting CP. 1997. Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Sci*. 6:464–468.
Ponting CP, Phillips C, Davies KE, Blake DJ. 1997. PDZ domains: targeting signalling molecules to sub-membranous sites. *Bioessays*. 19:469–479.
Putnam NH, Srivastava M, Hellsten U, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317:86–94.
Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of metazoa. *Mol Biol Evol*. 25:664–672.
Sakarya O, Armstrong KA, Adamska M, Adamski M, Wang IF, Tidor B, Degnan BM, Oakley TH, Kosik KS. 2007. A post-synaptic scaffold at the origin of the animal kingdom. *PLoS ONE*. 2:e506.
Sakarya O, Kosik KS, Oakley TH. 2008. Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony. *Bioinformatics* 24:606–612.
Schultz J, Hoffmuller U, Krause G, Ashurst J, Macias MJ, Schmieder P, Schneider-Mergener J, Oschkinat H. 1998. Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels. *Nat Struct Biol*. 5:19–24.
Sheng M, Sala C. 2001. PDZ domains and the organization of supramolecular complexes. *Annu Rev Neurosci*. 24:1–29.
Smith JM. 1970. Natural selection and the concept of a protein space. *Nature* 225:563–564.
Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, Chishti AH, Crompton A, Chan AC, Anderson JM, Cantley LC. 1997. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* 275:73–77.
Steuer R, Kurths J, Daub CO, Weise J, Selbig J. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18(Suppl 2):S231–S240.
te Velthuis AJ, Admiraal JF, Bagowski CP. 2007. Molecular evolution of the MAGUK family in metazoan genomes. *BMC Evol Biol*. 7:129.
te Velthuis AJ, Isogai T, Gerrits L, Bagowski CP. 2007. Insights into the molecular evolution of the PDZ/LIM family and identification of a novel conserved protein motif. *PLoS ONE*. 2:e189.
Tochio H, Zhang Q, Mandal P, Li M, Zhang M. 1999. Solution structure of the extended neuronal nitric oxide synthase PDZ domain complexed with an associated peptide. *Nat Struct Biol*. 6:417–421.
Tonikian R, Zhang Y, Sazinsky SL, et al. 2008. A specificity map for the PDZ domain family. *PLoS Biol*. 6:e239.
von Ossowski I, Oksanen E, von Ossowski L, Cai C, Sundberg M, Goldman A, Keinanen K. 2006. Crystal structure of the second PDZ domain of SAP97 in complex with a GluR-A C-terminal peptide. *FEBS J*. 273:5219–5229.
Wilken C, Kitzing K, Kurzbauer R, Ehrmann M, Clausen T. 2004. Crystal structure of the DegS stress sensor: how a PDZ domain recognizes misfolded protein and activates a protease. *Cell* 117:483–494.
Yeates DK. 2005. Groundplans and exemplars: paths to the tree of life. *Cladistics* 11:343–357.
Yeh RF, Lim LP, Burge CB. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res*. 11:803–816.
Zhang Y, Dasgupta J, Ma RZ, Banks L, Thomas M, Chen XS. 2007. Structures of a human papillomavirus (HPV) E6 polypeptide bound to MAGUK proteins: mechanisms of targeting tumor suppressors by a high-risk HPV oncoprotein. *J Virol*. 81:3618–3626.
Zhang Y, Yeh S, Appleton BA, et al. 2006. Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. *J Biol Chem*. 281:22299–22311.
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol*. 7:203–214.