Computational Biology and Chemistry 35 (2011) 24–33

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Bacterial protein structures reveal phylum dependent divergence

Matthew D. Shortridge [a], Thomas Triplet [b,1], Peter Revesz [b], Mark A. Griep [a], Robert Powers [a,*]

[a] Department of Chemistry, University of Nebraska-Lincoln, 722 Hamilton Hall, Lincoln, NE 68588-0304, United States
[b] Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0115, United States

ARTICLE INFO

Article history:
Received 27 July 2010
Received in revised form
28 December 2010
Accepted 29 December 2010

Keywords:
Proteins
Structure
Sequence
Function
Evolution

ABSTRACT

Protein sequence space is vast compared to protein fold space. This raises important questions about how structures adapt to evolutionary changes in protein sequences. A growing trend is to regard protein fold space as a continuum rather than a series of discrete structures. From this perspective, homologous protein structures within the same functional classification should reveal a constant rate of structural drift relative to sequence changes. The clusters of orthologous groups (COG) classification system was used to annotate homologous bacterial protein structures in the Protein Data Bank (PDB). The structures and sequences of proteins within each COG were compared against each other to establish their relatedness. As expected, the analysis demonstrates a sharp structural divergence between the bacterial phyla *Firmicutes* and *Proteobacteria*. Additionally, each COG had a distinct sequence/structure relationship, indicating that different evolutionary pressures affect the degree of structural divergence. However, our analysis also shows the relative drift rate between sequence identity and structure divergence remains constant.
© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Quantifiable models of protein evolution are useful for developing robust tools to identify suitable drug-binding sites, to predict increases in susceptibility to a human genetic disease, and to predict and modify organismal niches. Some of the strongest arguments in favor of biological evolution draw from studies on protein evolution using sequence homology (Do and Katoh, 2008). Multiple sequence alignments are routinely used to create phylogenetic relationships (Chang et al., 2008; Feng, 2007), which highlights sequence variability between organisms. The accepted view of protein evolution is that changes to the protein's gene sequence are selected and modulated by a number of factors that includes structure (Pal et al., 2006; Rocha, 2006).

What is the impact on protein structure as its sequence undergoes genetic drift? Maintaining the correct protein fold is fundamental to preserving its function (Forouhar et al., 2007), but evolving the sequence would also be expected to result in structural changes (Chothia and Lesk, 1986; Rost, 1999). The resulting observation is that sequence determines a protein's structure, but the structure is relatively invariant over a large range of sequences. This is highlighted by the tremendous difference between the number of known protein structures versus protein folds (Sadreyev and Grishin, 2006). Even though the Protein Data Bank (PDB) (Berman et al., 2000) contains 66,083 protein structures as of June 22, 2010, there are only 1233 unique topologies and 1195 unique folds in the CATH (Orengo et al., 1997) and SCOP (Murzin et al., 1995) structure classification databases, respectively. The significant reduction in the number of protein folds relative to the number of protein sequences implies a much stronger correlation between structure and function. Correspondingly, protein structures are generally viewed as more conserved relative to its sequence and recent studies have attempted to quantify this statement (Illergard et al., 2009).

The explicit reason for the reduction in fold space remains unclear. However, some have suggested that protein fold space may be more appropriately described as a continuum instead of a collection of discrete folds (Kolodny et al., 2006). In this manner, a protein fold should be considered as being plastic, where sequence changes are accommodated by local perturbations in the structure while maintaining the general characteristics of a particular fold (Illergard et al., 2009; Panchenko et al., 2005; Williams and Lovell, 2009). Correspondingly, the genetic drift in a protein's sequence may imply a similar gradual divergence in structure instead of a sudden dramatic transition to a new fold. From this perspective, a comparative analysis of homologous proteins should identify correlated rates of structure and sequence divergence. Previous studies have looked at

Abbreviations: FSS, fractional Structure Similarity; COG, Cluster of Orthologous Groups; PDB, Protein Data Bank; Split, clusters showing strong phylogenetic split pattern based on structure; Split + 1, clusters showing strong phylogenetic split pattern with one outlier based on structure; Starburst, clusters with variable phylogenetic patterns based on structure; Z_{AA} and Z_{BB}, Dali *Z*-scores for self comparisons; Z_{AB}, Dali *Z*-scores for pairwise comparisons; θ_{FSS}, structure similarity ratio; θ_{SeqID}, sequence similarity ratio.

* Corresponding author. Tel.: +1 402 472 3039; fax: +1 402 472 9402.
E-mail address: rpowers3@unl.edu (R. Powers).
[1] Present address: Centre for Structural and Functional Genomics, Concordia University, Montreal, QC, Canada H4B-1R6.

1476-9271/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2010.12.004

homologous structure similarity before but the datasets did not try to show structure divergence consequences on phylogenetic relationships (Illergard et al., 2009; Panchenko et al., 2005; Williams and Lovell, 2009). To help understand how protein plasticity affects organism divergence, we compared 48 sets of homologous protein families annotated in the COG database for two bacterial phyla, *Proteobacteria* and *Firmicutes*.

2. Materials and methods

2.1. COG assignment of the Protein Data Bank

Assignment of each bacterial protein in the PDB to a COG number in the clusters of orthologous groups (Tatusov et al., 2003) database required downloading the complete sequence lists from both databases and running a pairwise Basic Local Alignment Search Tool (BLAST) comparison. The pairwise protein BLAST search was run using the Protein Mapping and Comparison Tool (PROMPT v0.9.2) (Schmidt and Frishman, 2006) that allowed for large pairwise BLAST searching and reported the best match between the two databases. The BLAST search was run using the BLOSUM62 matrix with a gap penalty of 11, gap extension penalty of 1, a word size of 5, and a BLAST expectation threshold (*E*-value) of 10^{-9}. This *E*-value was used to unambiguously match genes in the COG database with proteins in the PDB. All PDB-to-COG matches were reported and stored in our PROFESS (**PRO**tein **F**unction, **E**volution, **S**tructure, and **S**equence) database (http://cse.unl.edu/~profess/, Triplet et al., 2010).

After matching structures to their representative COG, each PDB entry was matched with its source organism and phylum. The data set was then filtered according to the number of unique organisms. Specifically, only those COGs with structures from two or more different source organisms in both *Proteobacteria* and *Firmicutes* were analyzed further.

2.2. Pairwise structure comparison

The pairwise structure comparison program DaliLite v2.4.2 (Holm and Park, 2000) was installed on our 16-node Dual Athlon AMD 2.13 GHz with 1 GB of RAM Beowulf cluster running CentOS 4.4 Linux with a 2.25TB RAID array. A C-shell script matches the PDB files from each *Proteobacteria–Proteobacteria* comparison (−/−), *Firmicutes–Firmicutes* comparison (+/+) and *Proteobacteria–Firmicutes* comparison (−/+) and then submits the job to the program DaliLite. Each structural comparison takes approximately 2–10 min, depending on the size and relative similarity of structures. The total time to run all 63,504 comparisons was approximately 7 weeks.

The shell script extracts all structural comparison information reported by DaliLite (comparison files, rmsd, %Sequence ID, *Z*-score) on a per chain basis. A single PDB file may contain multiple protein chains, where each chain may have a separate COG assignment. All structure information is stored in our PROFESS database (Triplet et al., 2010), which is parsed to find the largest *Z*-score for each pairwise structure comparison. The largest *Z*-score represents the best structure comparison for a pair of proteins and ensures that the correct PDB chains were used for the analysis and the correct COG assignments were made. All best matches from each COG were used to calculate the Fractional Structure Similarity score (FSS) described by Eq. (1).

$$\mathrm{FSS} = \frac{Z_{AB}}{\max(Z_{AA}, Z_{BB})} \quad (1)$$

where Z_{AB} was the *Z*-score for comparing proteins A and B, Z_{AA} was the *Z*-score when protein A was compared to itself and Z_{BB} was the *Z*-score when protein B was compared to itself. Thus, Z_{AA} and Z_{BB} represent the *Z*-score that can be achieved for perfect similarity.

2.3. Manual filtering and data analysis

Manual refinement of the dataset included verification of each PDB assignment to a COG and filtering out redundantly solved structures from the same organism. When multiple structures were reported from the same organism (or organism with synonymous name), the structure that gave the largest Dali *Z*-score within the COG was kept while remaining structures were discarded from the analysis. This confirmed a single best PDB-COG match for each organism. Manual refinement was accomplished by opening all PDB IDs within a COG and checking biological information against the PDB (http://www.rcsb.org/pdb/home), COG (http://www.ncbi.nlm.nih.gov/COG/) and the NCBI (http://www.ncbi.nlm.nih.gov/) web servers. Consistency in functional and structural assignment within a COG coupled with low *E*-values between COG and PDB confirmed the best matches were functionally the same protein. Additionally, manual refinement was used to verify uniform sample conditions (i.e., the same ligand bound to all proteins within a COG or all proteins correspond to wild-type sequences) for cases of redundantly solved structures. The PDB to CATH linkage was obtained directly from the CATH v3.2 database. The CATH classification for each structural domain for the PDB files listed in Table 1 was manually verified using the CATH search engine. This was important because, even though 32 of the 48 COG structure families are single-domain proteins, the remaining 16 COG structure families have two or three domains.

2.4. Structure based phylogenetic trees

In addition to pairwise alignment, all the protein structures from each COG were simultaneously aligned using the multiple structure alignment program MAMMOTH-multi (http://ub.cbm.uam.es/mammoth/mult/) (Lupyan et al., 2005). The resulting aligned structures and the structure-based sequence alignment was used with in-house software to calculate an all-versus-all matrix of per-residue Cα distances. Standard bootstrapping techniques was then applied to the all-versus-all matrix of per-residue Cα distances to generate 100 distance-matrix tables. Columns of structure-based sequence alignments with the corresponding Cα distances were randomly selected until the total number of columns in the original sequence alignment was reached. The resulting set of Cα distances was then used to calculate a root mean square deviation (rmsd) between each pair of structures in the matrix. The 100 distance-matrix tables were imported into PHYLIP v3.68 (Felsenstein, 1989) to generate a consensus phylogenetic tree with bootstrap confidence levels (Efron et al., 1996).

Each set of 100 bootstrapped distance matrices was analyzed by the Fitch–Margoliash method implemented in PHYLIP. Each matrix was jumbled with 100 replicates using a random number generator seed. This resulted in 10,000 unique and random distance matrices for each COG. The best tree was identified with the program Consense implemented in PHYLIP using the extended majority rule conservation. Since the bootstrapped trees do not show distance relationship, the original distance matrix generated by MAMMOTH-multi was used to generate a distance based phylogenetic tree. Each original distance matrix was jumbled with 100 replicates using a random number seed. The distance based phylogenetic tree was drawn using the program Drawtree implemented in PHYLIP. Each tree was visually inspected and compared with the DaliLite analysis using the bootstrap values to determine if a tree fit the starburst, split or split + 1 classification.

Table 1
COG structure families.[a]

COG	Function	$S_{go_sim}(COG)$[b]	Phylogenetic structure tree[c]	CATH	Which domain?
28	Thiamine pyrophosphate requiring enzymes	0.59	Split	3.40.50.970	1st
				3.40.50.1220	2nd
				3.40.50.970	3rd
39	Malate/lactate dehydrogenases	0.80	Split	3.40.50.720	Single domain
394	Protein-tyrosine-phosphatase	0.61	Split	3.40.50.270	Single domain
446	Uncharacterized NAD (FAD)-dependent dehydrogenases	0.85	Split	3.50.50.60	1st
				3.50.50.60	2nd
				3.30.390.30	3rd
604	NADPH:quinone reductase and related Zn-dependent oxidoreductases	0.88	Split	3.40.50.720	Single domain
605	Superoxide dismutase	0.76	Split	[d]	1st and 2nd
742	N6-adenine-specific methylase	0.73	Split	[d]	Single domain
813	Purine-nucleoside phosphorylase	0.87	Split	3.40.50.1580	Single domain
1012	NAD-dependent aldehyde dehydrogenases	0.58	Split	3.40.309.10	1st and 2nd
1057	Nicotinic acid mononucleotide adenylyltransferase	0.95	Split	3.40.50.620	Single domain
1075	Predicted acetyltransferases and hydrolases with the alpha/beta hydrolase fold	0.70	Split	3.40.50.1820	Single domain
1607	Acyl-CoA hydrolase	0.87	Split	[d]	Single domain
1940	Transcriptional regulator/sugar kinase	0.31	Split	3.30.420.40	1st
				3.30.420.160	2nd
2124	Cytochrome P450	0.80	Split	1.10.630.10	Single domain
2188	Transcriptional regulators	0.89	Split	3.40.1410.10	Single domain
242	N-formylmethionyl-tRNA deformylase	0.87	Split with HGT	3.90.45.10	Single domain
1052	Lactate dehydrogenase and related dehydrogenases	0.89	Split with HGT	3.40.50.720	1st and 2nd
2141	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases	0.76	Split with HGT	3.20.20.30	Single domain
3832	Uncharacterized conserved protein	1.00	Split with HGT	3.30.530.20	Single domain
110	Acetyltransferase (isoleucine patch superfamily)	0.56	Starburst	2.160.10.10	Single domain
171	NAD synthase	0.85	Starburst	3.40.50.620	Single domain
251	Putative translation initiation inhibitor, yjgF family	0.00	Starburst	3.30.1330.40	Single domain
346	Lactoylglutathione lyase and related lyases	0.11	Starburst	3.10.180.10	Single domain
366	Glycosidases	0.51	Starburst	3.20.20.80	1st
				3.90.400.10	2nd
				2.60.40.1180	3rd
454	Histone acetyltransferase HPA2 and related acetyltransferases	0.83	Starburst	3.40.630.30	Single domain
491	Zn-dependent hydrolases, including glyoxylases	0.50	Starburst	3.60.15.10	Single domain
500	SAM-dependent methyltransferases	0.59	Starburst	3.40.1630.10	1st
				3.40.50.150	2nd
526	Thiol-disulfide isomerase and thioredoxins	0.96	Starburst	3.40.30.10	Single domain
590	Cytosine/adenosine deaminases	0.70	Starburst	3.40.140.10	Single domain
637	Predicted phosphatase/phosphohexomutase or 1.10.164.10	0.52	Starburst	3.40.50.1000	1st
				1.10.150.240	2nd
664	cAMP-binding proteins	0.50	Starburst	2.60.120.10	1st
				1.10.10.10	2nd
745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.73	Starburst	3.40.50.2300	Single domain
753	Catalase	0.93	Starburst	[d]	Single domain
778	Nitroreductase	0.64	Starburst	3.40.109.10	Single domain
784	FOG: CheY-like receiver	0.48	Starburst	3.40.50.2300	Single domain
796	Glutamate racemase	0.92	Starburst	3.40.50.1860	1st and 2nd
1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	0.84	Starburst	3.40.50.720	Single domain
1151	6Fe–6S prismane cluster-containing protein	0.71	Starburst	3.40.50.2030	1st
				3.40.50.2030	2nd
				1.20.1270.30	3rd
1309	Transcriptional regulator	0.80	Starburst	1.10.10.60	1st
				1.10.357.10	2nd
1396	Predicted transcriptional regulators	0.54	Starburst	1.10.260.40	1st
				2.60.120.10	2nd
1404	Subtilisin-like serine proteases	0.60	Starburst	3.40.50.200	Single domain
1733	Predicted transcriptional regulators	1.00	Starburst	[d]	Single domain
1846	Transcriptional regulators	0.85	Starburst	1.10.10.10	Single domain
2159	Predicted metal-dependent hydrolase of the TIM-barrel fold	0.83	Starburst	3.20.20.140	Single domain
2367	Beta-lactamase class A	0.93	Starburst	3.40.710.10	Single domain
2730	Endoglucanase	0.88	Starburst	3.20.20.80	Single domain
3693	Beta-1,4-xylanase	0.89	Starburst	3.20.20.80	Single domain
4948	L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily	0.71	Starburst	3.30.390.10	1st
				3.20.20.120	2nd

[a] COG structure families have two or more represented structures from among the *Firmicutes* and two or more from among the *Proteobacteria*
[b] Normalized GO functional similarity score between each protein's GO term set and the consensus GO term set for the COG (Eq. (2))
[c] "Split" means the *Firmicutes* and *Proteobacteria* proteins were strongly separated from one another, "Starburst" means there was little to no evidence for a split according to phyla, and "Split with HGT" means there was strong evidence for a split according to phyla with the exception of one protein, which may indicate horizontal gene transfer (described as split + 1 in the text). See Supplementary Table IS for a list of the PDB files associated with each COG.
[d] The protein structures in this COG family are in the CATH holding pen awaiting manual domain separation and/or final CATH assignment.

2.5. Measuring functional similarity within a COG

Each protein in our dataset was annotated with the corresponding Gene Ontology (Ashburner et al., 2000) identification number found in the PDB. By definition, a strong consensus requires each protein to share the same GO term. Instead, a weak consensus set of GO terms was generated for each COG, where only a majority of proteins are required to share the same GO term. A distance was measured between the weak consensus set and the set of GO terms assigned to each individual protein. An average, normalized distance is reported for each COG, where a score of 1 indicates an identical functional classification and a score of 0 indicates a lack of functional similarity. The normalized GO functional similarity score between each protein's GO term set and the consensus GO term set for the COG was measured as follows:

$$S_{go_sim}(COG) = \sum_{p \in COG} \frac{\left|GO_{cog}(p) \cap GO_{cog_wc}\right|}{\left|GO_{cog}(p) \cup GO_{cog_wc}\right|} \quad (2)$$

where $S_{go_sim}(COG)$ is the normalized GO functional similarity score, GO_{cog_wc} denotes the weak consensus set of GO terms for the COG, and $GO_{cog}(p)$ denotes the set of GO terms set for each protein p in the COG.

3. Results

3.1. Creating the COG structure families

Current functional annotation tools available in the PDB include the Gene Ontology (GO) (Ashburner et al., 2000) and Enzyme Classification (EC) (Schomburg et al., 2004). Unfortunately, due to the potential for convergence of function, these annotation tools are not useful for the study of homologous structures. To accurately observe phylum dependent structure divergence of proteins, it is important to construct a dataset of functionally similar orthologs. Among the 20 resources for structural classification of proteins, the clusters of orthologous groups (COGs) scheme is the only one that attempts to identify orthology (Ouzounis et al., 2003) while providing moderate functional information. Therefore, each sequence and structure in the PDB was annotated with one COG number. Additionally, each protein was annotated with GO numbers and the relative functional similarity for each COG was measured (Table 1). This was achieved by developing the PROFESS database (Triplet et al., 2010) that contains the PDB to COG annotations among other biologically relevant information. This includes associating each structure with its phyla classification, which allowed for the structures from *Firmicutes* and *Proteobacteria* to be easily selected for further analysis (Table S1).

The most recent COG database was created by finding the genome-specific best-hit for each gene in 66 unicellular genomes (50 bacteria, 13 archaea, and 3 eukaryota). Specifically, the orthologs present in three or more genomes were detected automatically and then multidomain proteins were manually split into component domains to eliminate artifactual lumping. The online COG database contains 192,987 sequences distributed among 4876 COGs, accounting for 75% of genes in these 66 genomes.

At the time of our COG-to-PDB annotation, the PDB included 45,368 protein structures, although many of them were composed of multiple subunits (and therefore associated with an even larger number of sequences). The two best-represented bacterial phyla, which accounts for nearly one-fourth of all structures in the PDB, were selected for annotation. The PDB contains 8298 *Proteobacteria* protein structures and 3416 *Firmicutes* structures. The sequences for each of these structures were compared to the COG reference sequences using the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). An expectation cut-off of 1×10^{-9} was used to maximize the likelihood of matching each PDB with its correct COG. The BLAST comparison matched 82% of the *Firmicutes* and *Proteobacteria* sequences to specific COGs, resulting in the clustering of 2728 *Firmicutes* structures and 6881 *Proteobacteria* structures. Of these hits, 27% were 100% identical to the COG reference sequence and 97% matched with greater than 50% sequence identity. To carry out our comparative study, we selected only those COGs that contained a minimum of two *Firmicutes* organisms and two *Proteobacteria* organisms. This requirement gave 281 unique COGs with a total of 3047 bacterial proteins (1066 *Firmicutes* and 1981 *Proteobacteria*). In addition to COG clustering, the eggNOG (http://eggnog.embl.de) (Muller et al., 2010) and OMA databases (Schneider et al., 2007) (http://omabrowser.org) were also mined for generating orthologous sets of protein structures. However, there was no set of proteins that met our criteria of two structures per phylum per cluster.

To further support the COG–PDB clusters, the overall functional similarity for each COG was determined by measuring the average distance between the Gene Ontology annotations for each protein and a weak consensus list of GO annotations (Table 1). Overall each COG exhibited high functional similarity (0.72 ± 0.21) with 1 being functionally identical and 0 being functionally dissimilar. In addition to the high sequence and structure similarity within each COG, the GO functional similarity measure provides further support that the proteins have been properly annotated to the correct COG. Nevertheless, there are three apparent outliers; COG0251 (putative translation initiation inhibitor, yjgF family), COG0346 (lactoylglutathione lyase and related lyases) and COG1940 (transcriptional regulator/sugar kinase) have GO similarity scores of 0, 0.11 and 0.31, respectively. The low GO similarity scores for these COGs are driven by the inclusion of unannotated proteins in the dataset. All six single-domain proteins associated with COG0251 are classified as a conserved hypothetical protein and have no associated GO terms. Of the seventeen single-domain proteins associated with COG0346, nine lack GO term assignments and have no functional annotation. Additionally for COG1940, two of the five two-domain proteins have no GO terms assigned to the structure.

3.2. Pairwise structure similarity

The pairwise structure comparison tool DaliLite (Holm and Park, 2000) was used to perform 63,504 pairwise comparisons between all of the proteins in our dataset. In total, the backbone structure similarity corresponded to 31,542 *Proteobacteria–Proteobacteria* comparisons (−/−), 12,674 *Firmicutes–Firmicutes* comparisons (+/+), and 19,288 *Proteobacteria–Firmicutes* comparisons (−/+). All comparisons were manually filtered within their respective COG to remove all but one redundantly solved structure (the largest contributor to the size reduction of the dataset), multiple or non-functionally relevant conformations (mutant protein, non-native experimental conditions, inhibited ligand complex), and the shorter of two protein structures. The final dataset contained 48 COGs (Table I) with a total of 1713 structural comparisons among 147 *Firmicutes* proteins from 58 unique organisms and 176 *Proteobacteria* proteins from 84 unique organisms.

The resulting Dali *Z*-scores from the pairwise structure comparisons were plotted against sequence identity (Fig. 1) to reveal a saturating relationship as the percent identity rose to 100%. The lowest observed *Z*-score was 5.7 with a corresponding 16% sequence identity. This *Z*-score was well above the minimum cutoff of 2.0 (dashed line) for matches that were two standard deviations above a random match. This lowest *Z*-score came from the comparison of two *Firmicutes* proteins in COG0346 (lactoylglutathione lyase and related lyases): 2QH0 (*Clostridium acetobutylicum*); and 2QQZ (*Bacillus anthracis*). The average *Z*-score for all comparisons

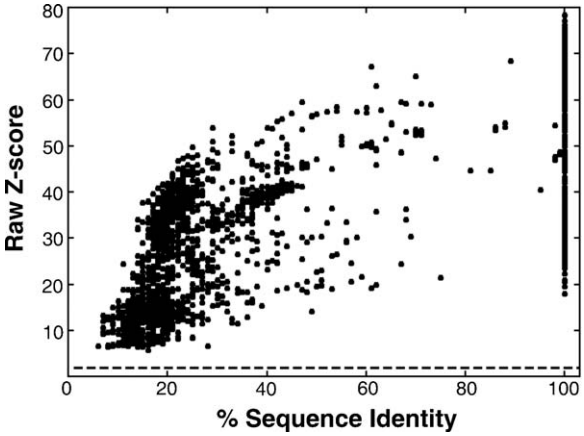

Fig. 1. The relationship between structure similarity and sequence identity for 48 COGs. Structure similarity is given as the raw *Z*-score, which increases as the protein length increases. The comparisons were for all proteins against all proteins, and include those for each protein against itself. The dashed line identifies a Dali *Z*-score of 2, which is the minimal limit for inferring structural similarity.

between these single-domain proteins was 27 ± 13, indicating that all structural matches were very significant even at sequence identities below 20%. All structure comparisons corresponding to 100% sequence identity in Fig. 1 result from a protein structure compared against itself. The inherent range in *Z*-scores at 100% sequence identity highlights the need to develop a normalized structure comparison score.

Since *Z*-scores increase as a function of the protein length, we normalized for this effect by calculating a Fractional Structure Similarity (FSS) scores (see Eq. (1)). When the pairwise FSS scores were plotted against sequence identity (Fig. 2), a hyperbolic curve was obtained with all FSS values below an upper-limit at each percent identity. In fact, 20% sequence identity yielded a maximal FSS of 60%. This FSS limit was observed when all of the data were used (Fig. 2A), when only the pairwise comparisons within either phyla were used (Fig. 2B and C), or when only the pairwise comparisons between the two phyla were used (Fig. 2D). The pairwise comparison plot between the two phyla (Fig. 2D) showed an abrupt cutoff at 61% sequence identity and a 0.84 FSS score. This was not an artifact created by culling the dataset, since a similar plot prior to the manual filtering also demonstrated the same effect (Supplemental Fig. 1).

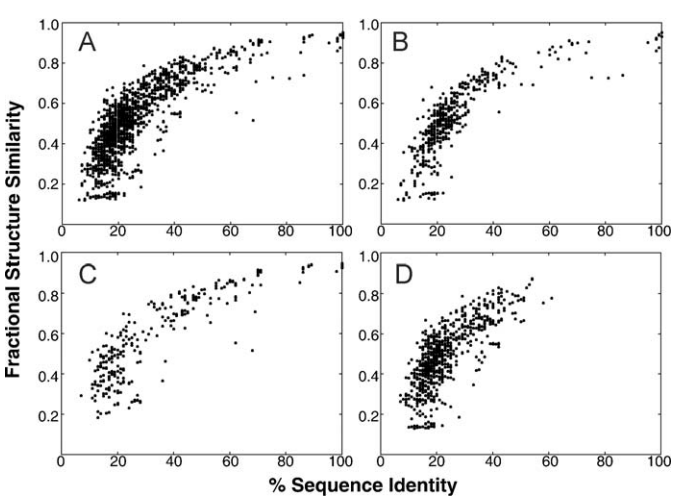

Fig. 2. The Fractional Structure Similarity (FSS) and sequence identity for 48 COGs. FSS was calculated using Eq. (1) to normalize the Dali *Z*-scores for their different sizes. The FSS values were plotted against sequence identity for (A) all the pairwise comparisons, (B) only *Proteobacteria–Proteobacteria* comparisons, (C) only *Firmicutes–Firmicutes* comparisons and (D) only *Proteobacteria–Firmicutes* comparisons.

The protein structures in COG0028 (thiamine pyrophosphate requiring enzymes) provides a useful example of the structural divergence that occurred after the *Firmicutes* and *Proteobacteria* phyla split. The overall fold is conserved between the phyla while there are discrete structural elements that are unique to each phylum. The two *Firmicutes* structures (Fig. 3A and B) yield a *Z* score of 59.6 and an FSS of 0.83, indicating very high structural conservation. There are more representative *Proteobacteria* structures that yield an average *Z*-score of 37.7 ± 1.6 and an average FSS of 0.58 ± 0.03. Again, the structures share a similar fold despite the slightly lower scores. Comparison of structures between the *Firmicutes* and *Proteobacteria* (Fig. 3C and D, respectively) phyla yield a lower *Z*-score of 34.8 ± 1.2 and a lower FSS of 0.49 ± 0.02 then the comparisons within each phylum. This suggests a divergence in structural details while conserving the overall fold. A detailed analysis reveals localized differences between the structures from the two phyla (see red highlights in Fig. 3C and D). In the *Firmicutes* representative structure, there is a continuous helix compared to helical breaks and loop insertions in the *Proteobacteria* structure. This is similar to the C-terminal domain of primase, where a long continuous helix found in the *Escherichia coli* structure is broken by a loop region in *B. stearothermophilus* (Bailey et al., 2007; Oakley et al., 2005; Su et al., 2006; Syson et al., 2005).

3.3. COG structure phylogenies

Structure based phylogenies were created from root-mean square differences (rmsd) in per residue Cα positions for optimally aligned protein structures using MAMMOTH-multi (Lupyan et al., 2005). A separate phylogenetic tree was generated for each COG, where three distinct patterns were observed (Table I): 15 trees exhibited a strong split at the phylum level, 29 exhibited a starburst pattern suggesting little to no evidence for a split according to phyla, and 4 exhibited a strong split at the phylum level but with the exception of a single structure (split + 1).

As shown in Table 1, the pattern of the structure based tree is not dependent on the relative GO functional similarity score for the proteins within each COG. All three tree patterns have a range of GO functional similarity scores with an average score of 0.75 ± 0.16, 0.88 ± 0.09 and 0.70 ± 0.24 for the split, split + 1, and starburst tree pattern, respectively. Overall the high GO similarity scores within each COG are high, indicating conserved and consistent functional annotations for each COG.

The 15 COG phylogenies with strong phylum-splitting patterns had two branches, one with closely related *Firmicutes* structures and the other with closely related *Proteobacteria* structures. Two examples are COG0028 (Thiamine pyrophosphate requiring enzymes) and COG0446 (Uncharacterized NAD/FAD-dependent dehydrogenases) (Fig. 4). The structures for both of these multi-domain COGs are classified in the CATH system as α/β 3-layer sandwiches, but differ in that COG0028 proteins have a Rossmann fold topology (Fig. 3) and COG0046 proteins have a FAD/NAD (P)-binding domain topology.

The 29 COGs with phylogenetic starburst patterns showed no evidence for the separation of structures according to phyla (Table 1). Two examples were COG0491 (Zn-dependent hydrolases) and COG1309 (transcriptional regulator) (Fig. 4). The CATH classification for COG0491 *Bacillus cereus* Zinc-dependent β-lactamase (PDB ID: 1BC2) (Fabiane et al., 1998) describes the protein as an α/β 4-layer sandwich with metallo-β-lactamase Chain A topology. The large category of β-lactamases constitutes a collection of enzymes that can be derived from any one of a group of proteins that bind, synthesize, or degrade peptidoglycans. The protein

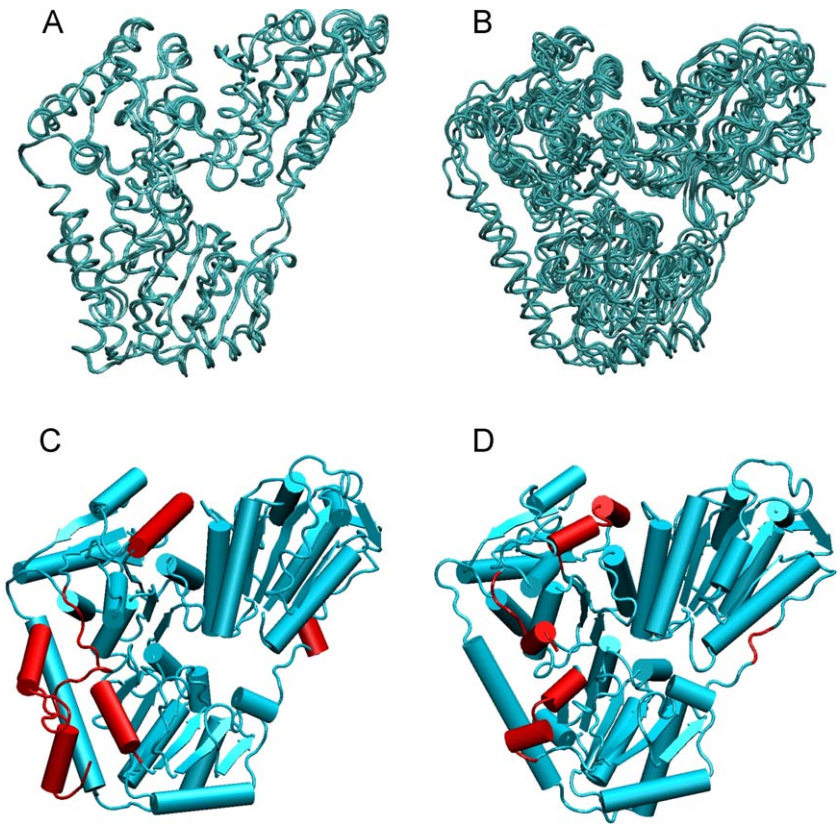

Fig. 3. Comparison of protein structures for COG0028 between two bacterial phyla. The protein structures for COG0028 thiamine pyrophosphate requiring enzymes show (A) that the two *Firmicutes* structures have highly overlapping structures and (B) that the four *Proteobacteria* structures are very similar to each another (see also the phylogenetic structure tree for COG0028 in Fig. 4). On the other hand, the major structural differences between the *Firmicutes* and *Proteobacteria* are highlighted in red on a representative *Firmicutes* (C) structure from *L. plantarum* (Lpl) (PDB ID: 1POW) (Muller et al., 1994) and the representative *Proteobacteria* structure (D) from *P. fluorescens* (Pfl) (PDB ID: 2AG0) (Mosbacher et al., 2005). (For interpretation of the references to color in text, the reader is referred to the web version of the article.)

structures assigned to COG0491 gave FSS scores with large standard deviations, as is evident from the separated clusters within the *Proteobacteria* arm of the phylogenetic tree.

The two-domain COG1309 structural family falls into one of two CATH topologies represented by Arc Repressor Mutant subunit A and Tetracycline Repressor domain 2. Only those structures similar to the Arc Repressor Mutant (subunit A) topology were used for the pairwise comparison, since it was the dominant fold in this COG. The protein structures in the COG1309 structure family gave low FSS scores. However, even with a low overall FSS, the average absolute *Z*-score was 13 ± 2 indicating that it has significant overall structure similarity. The high FSS deviations of the COG0491 structural family and the low average FSS scores of the COG1309 structural family both indicate rapid structural divergence following the phyla split, consistent with the observed starburst phylogenetic patterns.

Four COG structure phylogenies showed a strong split pattern with a single outlier (Fig. 5). This result provides further evidence for the observation of phyla split based on protein structure similarity. The presence of the outlier in a clear split pattern suggests either a horizontally transferred gene (Table I) or a potential paralog. For all four families [COG0242 (N-formylmethionyl-tRNA deformylase) COG1052 (lactate dehydrogenase and related dehydrogenases), COG2141 (coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases), and COG3832 (uncharacterized conserved protein)] there was a large Dali *Z*-score and reliable BLAST *E*-values, implying a correct match was made between COG and PDB. Additionally all four COGs exhibited high GO functional similarity scores suggesting a consistent functional assignment (Table 1). For COG0242, the *Bacillus cereus* gene *def* that encodes the N-formylmethionyl-tRNA deformylase protein (PDB ID: 1WS0) has been previously identified as a gene that has undergone horizontal gene transfer (Garcia-Vallve et al., 2000).

3.4. Structure divergence rates across phyla

As a way to quantify the relationship between structure and sequence differences, each phylogenetic tree was reduced to a single coordinate by calculating a structure similarity ratio (θ_{FSS}) and a sequence identity ratio (θ_{SeqID}). θ_{FSS} was determined for all 48 COGs by calculating an average FSS score for the *Proteobacteria–Firmicutes* structure comparisons, $\mathrm{Avg}(\mathrm{FSS}_{+/-})$, and dividing by the sum of the average *Proteobacteria–Proteobacteria*, $\mathrm{Avg}(\mathrm{FSS}_{-/-})$, and *Firmicutes–Firmicutes*, $\mathrm{Avg}(\mathrm{FSS}_{+/+})$, comparisons:

$$\theta_{\mathrm{FSS}} = \frac{\mathrm{Avg}(\mathrm{FSS}_{+/-})}{\mathrm{Avg}(\mathrm{FSS}_{+/+})/2 + \mathrm{Avg}(\mathrm{FSS}_{-/-})/2} \tag{3}$$

Similarly, a sequence identity ratio (θ_{SeqID}) was determined by calculating an average sequence identity for the *Proteobacteria–Firmicutes* structure comparisons, $\mathrm{Avg}(\mathrm{SeqID}_{+/-})$, and dividing by the sum of the average *Proteobacteria–Proteobacteria*, $\mathrm{Avg}(\mathrm{SeqID}_{-/-})$, and *Firmicutes–Firmicutes*, $\mathrm{Avg}(\mathrm{SeqID}_{+/+})$, comparisons:

$$\theta_{\mathrm{SeqID}} = \frac{\mathrm{Avg}(\mathrm{SeqID}_{+/-})}{\mathrm{Avg}(\mathrm{SeqID}_{+/+})/2 + \mathrm{Avg}(\mathrm{SeqID}_{-/-})/2} \tag{4}$$

In general, most starburst phylogenies (see representative COG0491 and COG1309 in Fig. 4) had a branch length between

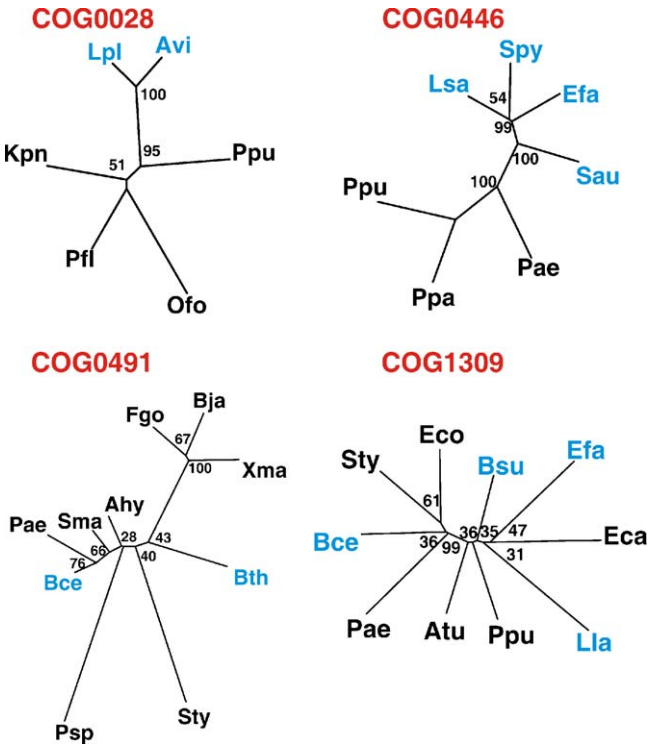

Fig. 4. Protein structure based phylogenetic trees highlighting the split and starburst patterns. The phylogenetic structure trees showed three different patterns: (top) strong split according to phyla; (bottom) starburst with no clear relationship to a common ancestor; and (Fig. 5) strong splits with the exception of one outlier. The *Firmicutes* protein structures are in blue and the *Proteobacteria* in black. The bootstrap values from 100 bootstrap replicates are indicated on branches and represent how often a branch appeared in the distance matrix. The two examples for the split pattern were from COG0028 (thiamine pyrophosphate requiring enzymes) and COG0446 (uncharacterized NAD(FAD)-dependent dehydrogenases). In the case of a strong split, the central branches were observed more than 95 times out of 100 replicate trials. The two examples for starburst pattern were from COG0491 (Zn-dependent hydrolases) and COG1309 (transcriptional regulator). For starburst patterns, very few branches were observed in more than two-thirds of the 100 replicate trials. The organism abbreviations are: *A. hydrophila* (Ahy); *A. tumefaciens* (Atu); *A. viridians* (Avi); *B. cereus* (Bce); *B. japonicum* (Bja); *B. subtilis* (Bsu); *B. thuriagienes* (Bth); *E. carotovora* (Eca); *E. coli* (Eco); *E. faecalis* (Efa); *F. gormanii* (Fgo); *K. pneumonia* (Kpn); *L. lactis* (Lla); *L. sanfranciscens* (Lsa); *L. plantarum* (Lpl); *O. formigens* (Ofo); *P. aeruginosa* (Pae); *P. fluorescens* (Pfl); *P. pantotrophus* (Ppa); *P. putida* (Ppu); *P. species* (Psp); *S. aureus* (Sau); *S. marcescens* (Sma); *S. typhimurium* (Sty); and *X. maltophilia* (Xma).

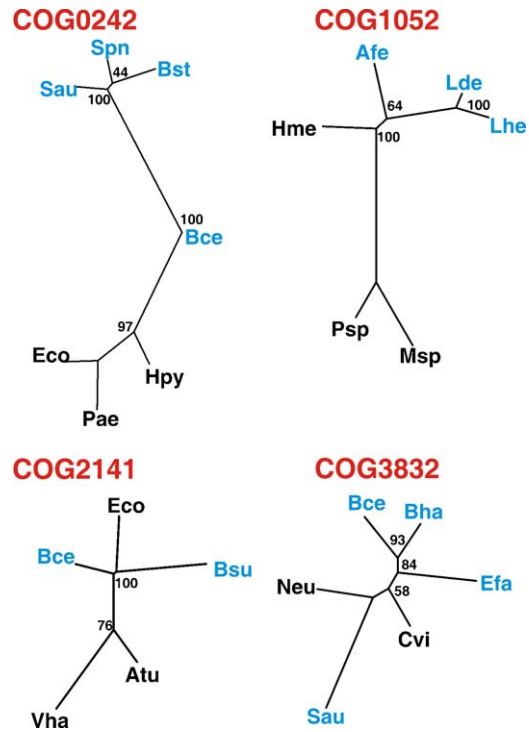

Fig. 5. Protein structure based phylogenetic trees highlighting the split + 1 pattern. Protein structure phylogenies of 4 COGs out of 48 had a strong split pattern with the exception of one outlier structure. The phylogenies were very reliable because the central branches were observed in 100 out of 100 replicate trials. When one *Firmicutes* or *Proteobacteria* protein structure clusters on a branch with the other phylum, its structure diverges from its closest relatives while resembling those of the other phyla. The COGs that fit this pattern are from COG0242 (N-formylmethionyl-tRNA deformylase), COG1052 (lactate dehydrogenase and related dehydrogenases), COG2141 (coenzyme F420-dependent N5, N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases), and COG3832 (uncharacterized conserved protein). The organism abbreviations are: *A. fermentans* (Afe); *A. tumefaciens* (Atu); *B. cereus* (Bce); *B. halodurans* (Bha); *B. stearothermophilus* (Bst); *B. subtilis* (Bsu); *C. violaceum* (Cvi); *E. coli* (Eco); *E. faecalis* (Efa); *H. methylovorum* (Hme); *H. pylori* (Hpy); *L. delbrueckii* (Lde); *L. helveticus* (Lhe); *M. species* (Msp); *N. europaea* (Neu); *P. aeruginosa* (Pae); *P. species* (Psp), *S. aureus* (Sau); *S. pneumoniae* (Spn); and *V. harveyi* (Vha).

members of different phyla that was much shorter than the branch lengths between members within the same phyla. That is, a starburst phylogeny was expected to have θ_{FSS} and θ_{SeqID} values greater than unity. Likewise, most split phylogenies had longer branches between phyla than within each phyla (see representative COG0028 and COG0446 in Fig. 4) and were expected to yield θ_{FSS} and θ_{SeqID} of less than unity.

When θ_{FSS} and θ_{SeqID} for all 48 COGs were plotted versus one another (Fig. 6), 79% of the starburst phylogenies were equal to or greater than unity for both structure and sequence whereas 84% of the split phylogenies were below a θ_{FSS} of 0.9 for structure and 73% of split phylogenies were below a θ_{SeqID} of 0.80 for sequence. This indicated that split phylogenies occur when the structure differences are less than their sequence differences. In addition, the plot of θ_{FSS} versus θ_{SeqID} conformed to a linear relationship regardless of the shape of the phylogenetic tree indicating that all homologous protein structure differences are constant with respect to homologous protein sequence differences ($\theta_{FSS} = 0.55\theta_{SeqID} + 0.45$; $R^2 = 0.7$). Thus, this curve represents the relative structural drift rate for each COG structural family between the two phyla. The slope indicates that structure branch

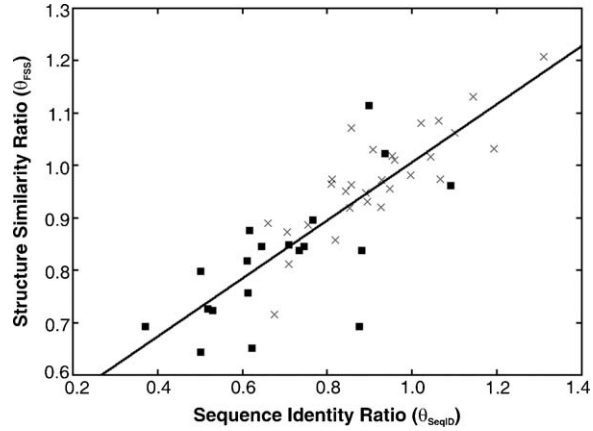

Fig. 6. Constant rate of structural drift. The relationship between structure and sequence change was constant regardless of the phylogenetic starburst (×) or split (■) pattern. Structure changes measured using a structure similarity ratio (θ_{FSS}), where the average FSS between members of the two phyla (*Firmicutes* versus *Proteobacteria*) was divided by the average FSS between members of the same phyla (see Eq. (3)). Sequence change was calculated similarly (see Eq. (4)). The best-fit line, $\theta_{FSS} = 0.55\theta_{SeqID} + 0.45$, yielded an R^2 of 0.70.

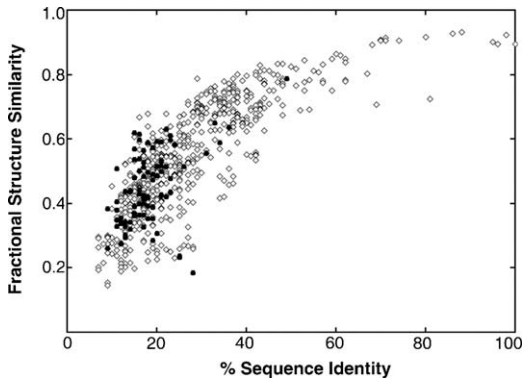

Fig. 7. Fold dependency on Fractional Structure Similarity (FSS) and sequence comparisons. The FSS between two CATH families, CATH 1.10 (●) CATH 3.40 (◊). CATH 1.10 (mainly α, orthogonal bundle) family is apparently limited to approximately 40% sequence identity and 0.6 FSS while CATH 3.40 (α/β, 3-layer (αβα) sandwich) fills in the complete curve. 87.5% of the COGs (7 of 8) represented by CATH 1.10 give a starburst structure similarity tree. Contrastingly, only 50% (12 of 24) of the COGs represented by CATH 3.40 give a starburst structure similarity tree. The remaining 12 COGs formed either split (11 of 12) or split + 1 (1 of 12).

lengths change approximately half as fast as sequence branch lengths.

3.5. Fold dependency on structure similarity

A plot of FSS versus sequence identity for the two most populated CATH families in our dataset (Fig. 7) was used to investigate if particular protein architectures are more amenable to structural changes. Thirty-one of 66 total domains (47%), the largest portion of our data set, are classified as CATH 3.40 (α/β, 3-layer (αβα) sandwich). The CATH 3.40 classification is more often associated with the split phylogenetic tree pattern (12 out of 22 total domains or 55%) then the starburst pattern (17 out of 39 total domains or 44%).

The second most populous CATH family is CATH 1.10 (mainly α, orthogonal bundle) with 11% of our COGs belonging to this CATH family. Most (85.7%) of the COGs (6 of 7) in the CATH 1.10 family are represented by the starburst phylogenetic tree pattern with only one COG represented by a split pattern. There appears to be a limit in structure similarity at approximately 0.6 FSS and a corresponding sequence identity limit at 40% for CATH 1.10 (Fig. 7, solid circles). This limit is not observed in the CATH 3.40 family (Fig. 7, open diamonds). The sequence and structure similarity limit for CATH 1.10 combined with a larger percentage of COGs assigned to the starburst family suggests that CATH 1.10 is more susceptible to mutations that affect the protein structure.

4. Discussion

There is an inherent challenge in obtaining an accurate functional annotation for a large set of proteins from a relatively small number of experimentally determined functions (Andrade, 2003; Frishman, 2007; Karp et al., 2001; Rentzsch and Orengo, 2009; Valencia, 2005). The available functional information is incomplete, ambiguous and error-prone (Benitez-Paez, 2009; Schnoes et al., 2009) and requires multiple sources (Rentzsch and Orengo, 2009) to improve the accuracy in the annotation of a protein. There is also the complicating factor of correctly distinguishing between orthologs and paralogs, where it has been previously noted that the COG database does include some paralog members (Dessimoz et al., 2006; Tatusov et al., 2003). Thus, the accuracy of our analysis of structural divergence is fundamentally dependent on a reliable functional assignment for each protein structure. Given these challenges, the independent and separate utilization of both COG and GO terms provides a reasonable and robust approach to identify clusters of functionally similar proteins. The overall high sequence (E-value $\leq 10^{-9}$, sequence identity $\geq$ 16%), structure (Z-score > 5.7) and GO term similarity (0.72 ± 0.21) within each COG supports this conclusion. The lack of identity for the GO term similarity scores should not be interpreted as evidence for functional divergence. GO terms are assigned based on a validated source. So, a missing GO term for a protein is more likely attributed to the fact that the protein has not been explicitly tested for the specified activity. Similarly, a protein being assigned a GO term does not provide definitive evidence that the function is relevant *in vivo* (Canevascini et al., 1996; Lindorff-Larsen et al., 2001; Otsuka et al., 2002; West et al., 2004).

The comparison of homologous protein structures with the same function provides quantitative evidence that protein structures diverged following the speciation events that created the modern bacterial phyla of *Firmicutes* and *Proteobacteria*. The abrupt cutoff at 61% sequence identity and 0.84 Fractional Structure Similarity observed between *Firmicutes* and *Proteobacteria* proteins was mirrored by an approximate 60% protein sequence identity between these two phyla observed by 16S rRNA sequence similarity (Konstantinidis and Tiedje, 2005a,b). Thus, this maximum observed sequence identity imparts limits to the maximum possible structure similarity between homologous proteins from these two phyla. This is consistent with prior observations that sequence identity $\leq$ 40–50% sometimes results in significant structural and functional differences (Chothia and Lesk, 1986; Rost, 1999; Rost, 2002). Furthermore, the results imply an inherent allowable structural plasticity that does not perturb function. Additionally, the random drift after speciation inexorably leads to non-identical structures despite maintenance of function. There are a number of cases where FSS was below 0.20 indicating a significant structural change. Proteins with completely different folds but the same function are extreme examples of the plasticity of the structure–function relationship and include such proteins as peptidyl-tRNA hydrolases (COG1990) (Powers et al., 2005), pantothenate kinase (KOG2201) (Yang et al., 2006), polypeptide release factors (Kisselev, 2002) and lysyl-tRNA synthetases (COG1190) (Ibba et al., 1997), these proteins are not in our dataset.

Forty percent of the COGs we examined have evolved slowly enough that it was possible to generate phylogenetic trees consistent with this ancient split. The other COGs have either evolved too rapidly or are otherwise subject to few evolutionary constraints to provide evidence for this split. This distinction between the COGs is clearly apparent from the comparison of θ_{FSS} and θ_{SeqID} in (Fig. 6). The slope of (Fig. 6) indicates a fixed relative structure drift rate, where structure changes half as fast as sequence across phyla. This correlation in the divergence of protein sequences and protein structures has additional ramifications beyond bacterial evolution. Our analysis implies a continuum of protein folds that adapt to large sequence changes by incurring local structural modifications (Illergard et al., 2009; Kolodny et al., 2006; Panchenko et al., 2005; Williams and Lovell, 2009). This continuum of protein folds makes it challenging to apply protein structural classification to identify function, as has been previously noted (Hadley and Jones, 1999; Pascual-Garcia et al., 2009).

Does the nature of the protein's three-dimensional structure play a role in protein structure divergence? Our analysis demonstrates that some proteins evolve slowly and maintain high sequence identity (>80%) and structure similarity (>0.80 FSS) while other proteins exhibit rapid evolution rates where sequence identity is $\leq$20% and FSS $\leq$ 0.40. This implies that the underlying architecture of a particular protein may be more or less amenable to amino-acid substitutions in order to maintain functional activity. A specific protein fold may have a higher intrinsic plasticity that enables it to readily accommodate sequence changes through local

conformational changes without a detrimental impact on activity. This is exactly what was observed.

Structural variations were localized to specific regions as illustrated by the comparison of the COG0028 protein structures (see Fig. 3). This is consistent with the observation that there are different structure divergence rates within a protein (Chirpich, 1975; Lin et al., 2007). Regions of the protein that do not impact biological activity are expected to yield a higher divergence rate and incur larger local structural changes (Chothia and Lesk, 1986; Lesk and Chothia, 1980). As a result, a fold with a relatively high plasticity would experience an elevated structural diversity between phyla, where the rate of change may closely parallel the mutation rate (Illergard et al., 2009). Conversely, another fold may be extremely sensitive to amino-acid substitutions, where minor sequence perturbations may result in a decrease in structural integrity and a corresponding loss of activity. This analysis is consistent with the known range of protein thermodynamic stabilities (Robertson and Murphy, 1997), and the general observation that most mutations destabilize protein structures (Sanchez et al., 2006).

For instance, CATH 1.10 was the second most abundant protein architecture observed in our study, comprising 11% of the total domains. It was very strongly associated with the fastest evolving protein structure and corresponds to an orthogonal α-helical bundle. Conversely, the highly populated CATH 3.40 is a 3-layer (αβα) sandwich with a slower evolution rate compared to CATH 1.10 structures. β-Sheets are strongly influenced by long-range interactions and, on average, have a higher hydrophobic environment compared to α-helices (Gromiha and Ponnuswamy, 1995). Effectively, the protein environment is an important factor in β-sheet folding (Parisien and Major, 2007). Since the stability and structure of a protein is strongly dependent on the integrity of the hydrophobic core (Vlassi et al., 1999), which is formed by the β-sheet in the 3-layer (αβα) sandwich, mutations in the β-sheet are probably less tolerated.

Our study illustrates the inherent value in solving structures for functionally identical proteins from multiple organisms. A major challenge in creating our COG-to-PDB dataset was the fundamental requirement to have structures from at least two *Firmicutes* organisms and two *Proteobacteria* organisms. Only 48 (∼1%) of the 4876 COGs meet this stringent requirement. The limited number of multiple homologous structures has partly occurred because structural biology efforts are focused on obtaining single representative structures for each functional class or protein fold (Chandonia and Brenner, 2005) and understandably biased toward therapeutically relevant proteins (Mestres, 2005). If we are to achieve a more accurate understanding of the relationship between the evolution of protein fold, protein sequence, and the organisms in which they function, the fields of bioinformatics and structural biology must expand their focus to include efforts to obtain a more diverse set of homologous protein structures.

Disclaimer

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases.

Acknowledgements

We would like to thank Venkat Ram Santosh from the University of Nebraska-Lincoln for his contribution to the GO functional similarity scores. This work was supported in part from the National Institute of Allergy and Infectious Diseases (Grant No. R21AI081154), from the Nebraska Tobacco Settlement Biomedical Research Development Funds, and a Nebraska Research Council Interdisciplinary Research Grant to R.P. The research was performed in facilities renovated with support from NIH (RR015468-01).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2010.12.004.

References

Altschul, S.F., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Andrade, M.A., 2003. Automatic Genome Annotation and the Status of Sequence Databases. Horizon Scientific Press, 107–121.
Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.
Bailey, S., et al., 2007. Structure of hexameric DnaB helicase and its complex with a domain of DnaG primase. Science 318, 459–463.
Benitez-Paez, A., 2009. Considerations to improve functional annotations in biological databases. OMICS 13, 527–532.
Berman, H.M., et al., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242.
Canevascini, S., et al., 1996. Tissue-specific expression and promoter analysis of the tobacco Itp1 gene. Plant Physiol. 112, 513–524.
Chandonia, J.M., Brenner, S.E., 2005. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins 58, 166–179.
Chang, G.S., et al., 2008. Phylogenetic profiles reveal evolutionary relationships within the "twilight zone" of sequence similarity. Proc. Natl. Acad. Sci. U.S.A. 105, 13474–13479.
Chirpich, T.P., 1975. Rates of protein evolution. Function of amino acid composition. Science 188, 1022–1023.
Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826.
Dessimoz, C., et al., 2006. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res. 34, 3309–3316.
Do, C.B., Katoh, K., 2008. Protein multiple sequence alignment. Methods Mol. Biol. (Totowa, NJ, USA) 484, 379–413.
Efron, B., et al., 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. U.S.A. 93, 7085–7090.
Fabiane, S.M., et al., 1998. Crystal structure of the zinc-dependent beta-lactamase from Bacillus cereus at 1.9 Å resolution: binuclear active site with features of a mononuclear enzyme. Biochemistry 37, 12404–12411.
Felsenstein, J., 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166.
Feng, J.-A., 2007. Improving pairwise sequence alignment between distantly related proteins. Methods Mol. Biol. (Totowa, NJ, USA) 395, 255–268.
Forouhar, F., et al., 2007. Functional insights from structural genomics. J. Struct. Funct. Genomics 8, 37–44.
Frishman, D., 2007. Protein annotation at genomic scale: the current status. Chem. Rev. 107, 3448–3466.
Garcia-Vallve, S., et al., 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10, 1719–1725.
Gromiha, M.M., Ponnuswamy, P.K., 1995. Prediction of protein secondary structures from their hydrophobic characteristics. Int. J. Pept. Protein Res. 45, 225–240.
Hadley, C., Jones, D.T., 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 7, 1099–1112.
Holm, L., Park, J., 2000. DaliLite workbench for protein structure comparison. Bioinformatics 16, 566–567.
Ibba, M., et al., 1997. A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. Science 278, 1119–1122.
Illergard, K., et al., 2009. Structure is three to ten times more conserved than sequence – a study of structural response in protein cores. Proteins 77, 499–508.
Karp, P.D., et al., 2001. Database verification studies of SWISS-PROT and GenBank. Bioinformatics 17, 526–532.
Kisselev, L., 2002. Polypeptide release factors in prokaryotes and eukaryotes: same function, different structure. Structure 10, 8–9.
Kolodny, R., et al., 2006. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Curr. Opin. Struct. Biol. 16, 393–398.
Konstantinidis, K.T., Tiedje, J.M., 2005a. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 102, 2567–2572.
Konstantinidis, K.T., Tiedje, J.M., 2005b. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187, 6258–6264.
Lesk, A.M., Chothia, C., 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J. Mol. Biol. 136, 225–270.
Lin, Y.-S., et al., 2007. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol. Biol. Evol. 24, 1005–1011.
Lindorff-Larsen, K., et al., 2001. Barley lipid transfer protein, LTP1, contains a new type of lipid-like post-translational modification. J. Biol. Chem. 276, 33547–33553.
Lupyan, D., et al., 2005. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 21, 3255–3263.

Mestres, J., 2005. Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery. Drug Discovery Today 10, 1629–1637.
Mosbacher, T.G., et al., 2005. Structure and mechanism of the ThDP-dependent benzaldehyde lyase from *Pseudomonas fluorescens*. FEBS J. 272, 6067–6076.
Muller, Y.A., et al., 1994. The refined structures of a stabilized mutant and of wild-type pyruvate oxidase from *Lactobacillus plantarum*. J. Mol. Biol. 237, 315–335.
Muller, J., et al., 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res. 38, D190–D195.
Murzin, A.G., et al., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Oakley, A.J., et al., 2005. Crystal and solution structures of the helicase-binding domain of *Escherichia coli* primase. J. Biol. Chem. 280, 11495–11504.
Orengo, C.A., et al., 1997. CATH – a hierarchic classification of protein domain structures. Structure 5, 1093–1108.
Otsuka, T., et al., 2002. CCl4-induced acute liver injury in mice is inhibited by hepatocyte growth factor overexpression but stimulated by NK2 overexpression. FEBS Lett. 532, 391–395.
Ouzounis, C.A., et al., 2003. Classification schemes for protein structure and function. Nat. Rev. Genet. 4, 508–519.
Pal, C., et al., 2006. An integrated view of protein evolution. Nat. Rev. Genet. 7, 337–348.
Panchenko, A.R., et al., 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. Proteins 61, 535–544.
Parisien, M., Major, F., 2007. Ranking the factors that contribute to protein beta-sheet folding. Proteins 68, 824–829.
Pascual-Garcia, A., 2009. Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. PLoS Comput. Biol.V 5.
Powers, R., et al., 2005. Solution structure of *Archaeglobus fulgidis* peptidyl-tRNA hydrolase (Pth2) provides evidence for an extensive conserved family of Pth2 enzymes in archea, bacteria, and eukaryotes. Protein Sci. 14, 2849–2861.
Rentzsch, R., Orengo, C.A., 2009. Protein function prediction – the power of multiplicity. Trends Biotechnol. 27, 210–219.
Robertson, A.D., Murphy, K.P., 1997. Protein structure and the energetics of protein stability. Chem. Rev. 97, 1251–1268.
Rocha, E.P., 2006. The quest for the universals of protein evolution. Trends Genet. 22, 412–416.
Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng. 12, 85–94.
Rost, B., 2002. Enzyme function less conserved than anticipated. J. Mol. Biol. 318, 595–608.
Sadreyev, R.I., Grishin, N.V., 2006. Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. BMC Struct. Biol. 6, 6.
Sanchez, I.E., et al., 2006. Point mutations in protein globular domains: contributions from function, stability and misfolding. J. Mol. Biol. 363, 422–432.
Schmidt, T., Frishman, D., 2006. PROMPT: a protein mapping and comparison tool. BMC Bioinform. 7, 331.
Schneider, A., et al., 2007. OMA browser-exploring orthologous relations across 352 complete genomes. Bioinformatics 23, 2180–2182.
Schnoes, A.M., et al., 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. 5, e1000605.
Schomburg, I., et al., 2004. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. 32, D431–433.
Su, X.C., et al., 2006. Monomeric solution structure of the helicase-binding domain of *Escherichia coli* DnaG primase. FEBS J. 273, 4997–5009.
Syson, K., et al., 2005. Solution structure of the helicase-interaction domain of the primase DnaG: a model for helicase activation. Structure 13, 609–616.
Tatusov, R.L., et al., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinform. 4, 41.
Triplet, T., et al., 2010. PROFESS: a PROtein Function, Evolution, Structure and Sequence database, Database 2010, baq011.
Valencia, A., 2005. Automatic annotation of protein function. Curr. Opin. Struct. Biol. 15, 267–274.
Vlassi, M., et al., 1999. A correlation between the loss of hydrophobic core packing interactions and protein stability. J. Mol. Biol. 285, 817–827.
West, G., et al., 2004. Crystallization and X-ray analysis of bovine glycolipid transfer protein. Acta Crystallogr., Sect. D: Biol. Crystallogr. D60, 703–705.
Williams, S.G., Lovell, S.C., 2009. The effect of sequence evolution on protein structural divergence. Mol. Biol. Evol. 26, 1055–1065.
Yang, K., et al., 2006. Crystal structure of a type III pantothenate kinase: insight into the mechanism of an essential coenzyme A biosynthetic enzyme universally distributed in bacteria. J. Bacteriol. 188, 5532–5540.