Heliyon 6 (2020) e05867

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Visualizing the superfamily of metallo-β-lactamases through sequence similarity network neighborhood connectivity analysis

Javier M. González *

Instituto de Bionanotecnología del NOA (INBIONATEC), Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de Santiago del Estero (CONICET-UNSE), G4206XCP Santiago del Estero, Argentina

ARTICLE INFO

Keywords:
Metallo-lactamase
Protein superfamily
Tanglegram
Sequence similarity network
Neighborhood connectivity

ABSTRACT

Protein sequence similarity networks (SSNs) constitute a convenient approach to analyze large polypeptide sequence datasets, and have been successfully applied to study a number of protein families over the past decade. SSN analysis is herein combined with traditional cladistic and phenetic phylogenetic analysis (respectively based on multiple sequence alignments and all-against-all three-dimensional protein structure comparisons) in order to assist the ancestral reconstruction and integrative revision of the superfamily of metallo-β-lactamases (MBLs). It is shown that only 198 out of 15,292 representative nodes contain at least one experimentally obtained protein structure in the Protein Data Bank or a manually annotated SwissProt entry, that is to say, only 1.3 % of the superfamily has been functionally and/or structurally characterized. Besides, neighborhood connectivity coloring, which measures local network interconnectivity, is introduced for detection of protein families within SSN clusters. This approach provides a clear picture of how many families remain unexplored in the superfamily, while most MBL research is heavily biased towards a few families. Further research is suggested in order to determine the SSN topological properties, which will be instrumental for the improvement of automated sequence annotation methods.

1. Introduction

The metallo-β-lactamase (MBL) superfamily comprises an ancient group of proteins found in all domains of life, sharing a distinctive αββα fold with a histidine-rich motif for binding of transition metal ions. Such characteristic αββα domain uniquely places the metal binding site at the bottom of a wide groove that evolved to accommodate varied substrates. The name was coined after the first superfamily members to be characterized: a group of zinc-dependent hydrolases produced by bacteria resistant to β-lactam antibiotics. These zinc-β-lactamases (ZBLs) hydrolyze the amide bond present in all β-lactams and thus render them ineffective. The first X-ray crystallographic report of a ZBL was that of BcII from *Bacillus cereus* 569/H/9 [1]. Despite its low resolution, the atomic model disclosed the new αββα fold and a single Zn(II) ion bound to a three-histidine motif, resembling the active site of carbonic anhydrases. Thus, BcII and ZBLs in general were believed to use a single Zn(II) ion to activate a water molecule for hydrolysis, paralleling the mechanism by which carbonic anhydrases catalyze carbon dioxide hydration. This hypothesis was soon questioned when the structure of ZBL CcrA from *Bacteroides fragilis* was published, disclosing a bimetallic zinc center, with the second zinc being coordinated to nearby Asp, Cys and His residues [2]. Besides, the second zinc was later found in *B. cereus* ZBL too [3, 4, 5], starting a decade-long controversy regarding the role of each zinc ion. Later on, it was found that monometallic ZBLs are rather exceptional and the hydrolysis reaction generally requires two Zn(II) ions [6, 7].

A great diversity of proteins evolved in the MBL superfamily by combining catalytic MBL domains and substrate recognition domains in a modular fashion. Subtle changes in the metal coordinating residue networks expand this diversity by enabling the coordination of different transition metals, particularly Zn(II), Mn(II), and Fe(II)/Fe(III) (Figure 1). Early attempts to build a systematic classification of the MBL superfamily were conducted by L. Aravind [8], as some of the very first applications of the PSI-Blast algorithm [9], who showed that many proteins other than ZBLs comprise the characteristic fold and histidine-rich metal-binding motif of MBLs, mapping key residues onto the structure of *B. cereus* ZBL. These observations were updated in 2001 by Daiyasu *et al.*, when additional crystal structures of MBL superfamily members were available [10]. At present, more than a hundred proteins have been shown to

* Corresponding author.
E-mail address: javierg@conicet.gov.ar.

https://doi.org/10.1016/j.heliyon.2020.e05867
Received 11 October 2020; Received in revised form 19 November 2020; Accepted 23 December 2020

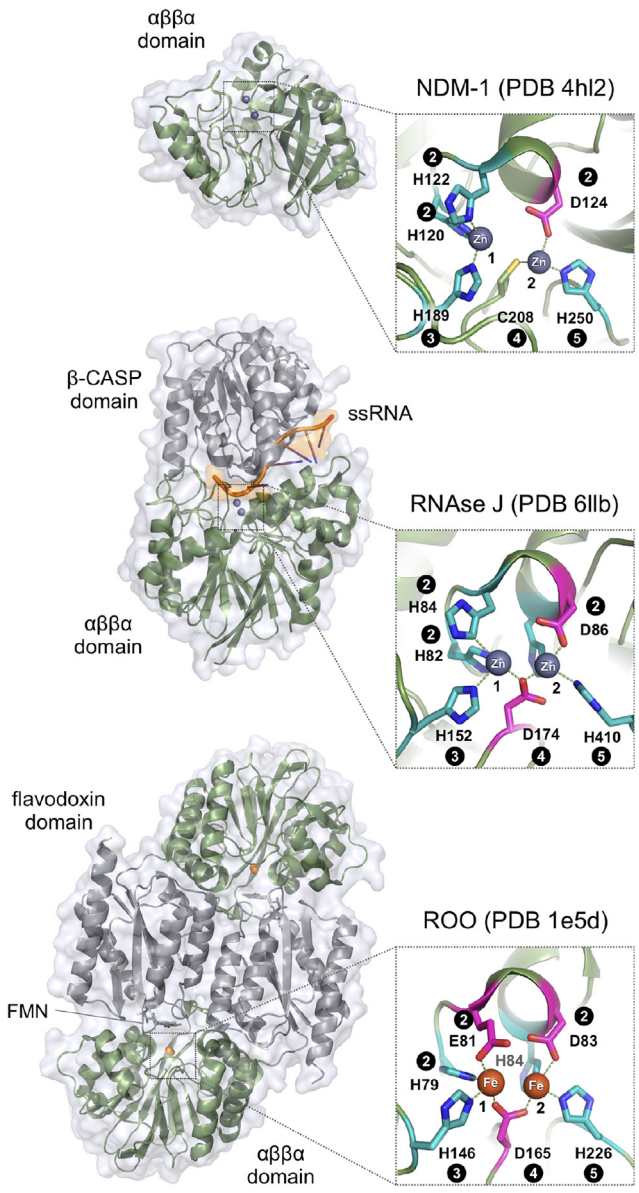


Figure 1. Structural diversity of MBLs. ZBLs like plasmid-borne *Klebsiella pneumoniae* NDM-1 (PDB 4hl2, *top*) comprise a single αββα domain (*green*), with the Zn(II) binding site at the bottom of an open groove, accessible to varied β-lactam antibiotics. B1 ZBLs exhibit unusual zinc ligands, including a cysteine residue, which is uncommon in catalytic Zn(II)-binding sites. Instead, habitual MBL metal ligand sets include only histidine and aspartic acid residues. For instance, RNAse J from *Methanolobus psychrophilus* (PDB 6llb, *middle*) comprises a phosphoesterase αββα domain and a β-CASP domain (*gray*) for single-stranded RNA binding (*orange*). Finally, MBL oxidoreductases such as the flavo-diiron protein ROO from *Desulfovibrio gigas* (PDB 1e5d, *bottom*) utilize non-heme Fe(II)/Fe(III) for catalysis, exhibiting a more acidic metal ligand set, in combination with an FMN-binding flavodoxin domain (*gray*), displaying a homodimeric quaternary structure. Metal ions are indicated as numbered spheres. Amino acid side chains follow the coloring scheme of Figure 2. Circled numbers indicate the corresponding motifs, as defined in Figure 2.

contain αββα domains through X-ray crystallography, whereas the InterPro 77.0 [11] database entry IPR001279 for the MBL superfamily includes about half a million members. Indeed, the MBL superfamily has grown astoundingly over the past 30 years, and an integrative revision is long overdue.

In recent years, protein families available in public databases have grown in number and size at unprecedented rates. Thus, improved methods for accurate analysis of large protein sequence datasets are urgently needed, since such a task is unattainable with the classical approach of multiple sequence alignment (MSA) plus phylogenetic tree calculation. A convenient approach introduced relatively recently by professor Babbitt group at UCSF is the construction of sequence similarity networks (SSN) [12]. SSNs comprise nodes representing a given set of polypeptide sequences interconnected with edges for a specified similarity cutoff value, and have been successfully applied to characterize a number of protein superfamilies in the past decade [13, 14, 15, 16, 17, 18, 19]. Nonetheless, identifying protein families within network clusters with missing experimentally-obtained functional or structural information is still an unsolved problem. Besides, the topological properties of SSN are largely unknown in comparison with classic models like random, small-world, and scale-free networks [20]. In this work, a large-scale MSA-based cladogram and a structure-based phenogram are calculated for the superfamily of metallo-β-lactamases in order to assist its phylogenetic reconstruction, providing a framework for an updated integrative revision. In addition, the neighborhood connectivity (NC) analysis [21] is introduced as an intuitive guide to search for uncharacterized new families within SSN clusters.

2. Materials and methods

2.1. Structural data harvesting and tanglegram calculation

All MBL protein sequences with available experimentally determined three-dimensional structure were retrieved from the Protein Data Bank (PDB) with the Dali Lite server [22], using structures PDB 2gmn and PDB 3i13 as queries. A set of 105 high-resolution structures was obtained after applying a 90 % sequence similarity cutoff. As well, an unrooted structural dendrogram was obtained for this set with the Dali Lite server all-against-all comparison tool, which calculates a distance matrix of *Z*-scores by aligning the structures all-against-all and outputs a dendrogram derived with the average linkage clustering method [23]. Next, the full amino acid sequence corresponding to each of these 105 structures were retrieved from the UniProt database [24], in order to avoid sequence artifacts like mutations and missing residues often found in PDB files. A structure-guided multiple sequence alignment (MSA) was calculated with Promals3D [25]. This MSA was manually edited with Jalview 2.9 [26] to discard highly gapped regions, by applying a 50 % alignment quality cutoff. The resulting MSA, comprising 105 sequences and 204 columns, was used to calculate a maximum likelihood cladogram with RAxML [27], running at the Cipres server [28]. A best-scoring bootstrapped tree was obtained after 1002 replicates, using the WAG substitution matrix as evolutionary model [29], and was displayed as a consensus cladogram by applying the 50 % majority rule. Finally, in order to compare the consensus sequence-based cladogram with the distance-based dendrogram topologies, a tanglegram matching corresponding taxa was calculated with the Neighbor Net Tanglegram algorithm [30], available in Dendroscope 3.5.9 [31], using the clade of B1&B2 zinc-β-lactamases as outgroup to root each tree. The tanglegram was adapted for display with FigTree 1.4.3 (available at http://tree.bio.ed.ac.uk/software/figtree/) and Corel Draw X7 (Corel). Protein structures were analyzed and graphically represented with PyMOL 1.8 (Schrödinger LLC).

2.2. Sequence data harvesting and SSN calculation

In order to prepare a representative sequence data sample of the MBL superfamily, the PF00753 Pfam database entry was selected as a starting point, which presently comprises 70,367 sequences (release Pfam 32.0, September 2018) [24]. The RP55 representative proteome MSA (62,213 sequences by 1,251 columns) was downloaded and manually edited with Jalview 2.9 [26], by removing truncated and misaligned sequences, highly gapped columns (more than 50 %); and deleting those sequences missing conserved positions corresponding to aspartic acid residues 29,

58, and 134 of human glyoxalase II, which was taken as a reference. The resulting MSA consisted of 55,076 sequences and 143 columns. Next, the full sequences present in this MSA set were retrieved from the UniProt 2019-10 database and reduced to a final set of 32,418 sequences, by applying a 70 % similarity cutoff with CD-Hit [32] and ensuring that all 105 sequences present in the tanglegram were included. A sequence similarity network (SSN) [12] was then calculated with this 32,418-sequence dataset, using the EFI-EST online tool [33]. The obtained representative node network comprised 15,292 nodes at 40 % sequence similarity, and 762,784 edges at 10–20 Blast pairwise similarity threshold. Topology network analysis was performed with NetworkAnalyzer 2.7 [34], as implemented in Cytoscape 3.7.1 [35]. Network statistics plots were prepared with SigmaPlot 12 (Systat Software). All figures were prepared with Corel Draw X7 (Corel).

3. Results and discussion

3.1. Unearthing ancestral relationships within the MBL superfamily

Tracing the evolutionary history of ancient protein superfamilies is often obscured by the inherent variability of amino acid sequences over long periods. Despite the divergence of primary structure, the three-dimensional fold of polypeptides is less sensitive to mutational events, retaining evolutionary information encoded in the arrangement of secondary structure elements. Thus, experimentally determined structures of proteins offer the possibility of common ancestry inference based on structural homology. Such *phenetic* methods are convenient for comparing proteins with similar folds but highly divergent amino acid sequences, in contrast to MSA-based *cladistic* methods, which are well suited to determine phylogenetic relationships between homologous proteins.

A structure-based approach for functional classification of MBLs was applied by Garau *et al.* in 2005, who used normalized root mean-square values as structural diversity estimates in order to calculate structure-guided phylogenies [36]. They conclude that structural similarity, as defined by differences in positions of Cα atoms of fitted homologous structures, is an acceptable estimate of evolutionary relatedness of proteins sharing comparable folds. A variant of this approach is herein employed, using the Dali *Z*-score as a more accurate estimate of structural similarity for a set of currently available experimental MBL structures. A distance matrix of Dali *Z*-scores comparing all-against-all full-length 105 selected MBL structures was used to construct the corresponding structural phenogram, that is, an unrooted tree whose branch lengths reflect structural similarity relationships between proteins, independently of their amino acid sequence. Next, the amino acid sequences of those 105 polypeptides were retrieved and aligned to construct a maximum-likelihood MSA-based bootstrapped unrooted consensus cladogram, whose topology reflects the sequence homology relationships between extant taxa according to a specific evolutionary model. Both dendrograms were then rooted using the B1&B2 ZBL clade as outgroup, since these enzymes are uniquely divergent MBLs due to their fast-evolving nature. The most distinctive feature of this outgroup is the presence of a Zn(II)-binding cysteine residue which is uncommon in catalytic Zn(II) sites, and has been shown to enable Zn(II) binding at limiting metal concentrations [7]. A tanglegram was then calculated with both trees, which consists of a graph of opposing dendrograms with lines connecting equivalent or corresponding taxa, rearranged so that the number of crossing connecting lines is minimal. This type of graph is widely used in Biology to illustrate processes like host-parasite, mutualistic, and symbiotic relationships, where both trees tend to comprise mirror images of each other, as a reflection of their shared topology and evolutionary history. Tanglegrams are used here to explore reciprocal similarities between structure and function of proteins. Since conserved structural features are substantiated by sequence adaptations to perform a specific function, sequence and structure can be assumed to evolve together, and should therefore give rise to dendrograms with the same topology. Crossing connectors between proteins would suggest that conserved residues typical of one group of proteins are found in a scaffold characteristic of different ones. Since the MSA consensus cladogram is not resolved at early nodes, a typical feature of phylogenies of divergent protein families, both trees can be rearranged so that no crossing connecting lines are needed between taxa (Figure 2).

3.2. Phenetic and cladistic considerations shed light on mutual MBL ancestors

ZBLs comprise a divergent polyphyletic group of MBLs, including subclasses B1, B2, and B3 [37]. It is important to note that, while ZBLs hydrolyze antibiotics by means of a metal-activated water molecule, most β-lactamases use a conserved serine residue in a completely different protein scaffold. In other words, the majority of β-lactamases are not metallic, and referring to ZBLs and MBLs in general simply as "β-lactamases" should be avoided, particularly when annotating these proteins in public databases. Besides, even though most members of the superfamily are devoid of β-lactamase activity, the acronym MBL has been adopted to annotate most members of the superfamily. The same convention is followed here to define any protein with at least one characteristic MBL domain, leaving the acronym ZBL to describe metallo-β-lactamases themselves.

As shown in the tanglegram and suggested previously [36, 38], B3 ZBLs form a phylogenetically distinct group as compared with B1&B2 enzymes, a clear example of how ZBL activity evolved twice within the superfamily. Motif 2 of B1, B2, and B3 ZBLs are characteristically of the form HxHxDX (where X is not a zinc ligand, typically Arg, Lys or small side chain residues), NxHxDR and HxHxDH, respectively. While B2 ZBLs are typically strict carbapenemases, B1 and B3 ZBLs display low substrate selectivity, and are able to hydrolyze all penicillins, cephalosporins and carbapenems of clinical use. Only monobactams remain insensitive to hydrolysis by ZBLs. Subclass B1 plasmid-borne ZBLs like IMP-1 (see Figure 2 for UniProt identifiers) became known in the '90s for their ability to hydrolyze carbapenems, the latest generation of β-lactam antibiotics available. 30 years later, pathogens expressing B1 enzymes like NDM-1 (Figure 1) still comprise one of the most cumbersome public health issues. In agreement with previous observations, B1&B2 enzymes are closely related and share a recent ancestor, along with a distinctive Zn(II)-binding cysteine at motif 4, supporting antibiotic resistance at limiting Zn(II) concentrations [7]. In contrast, B3 enzymes are typically chromosomal and replace this cysteine with residues unable to coordinate Zn(II) ions, like Ser, Ile, Val, Leu, and Met. In addition, all motif 2 histidines of B3 enzymes become zinc ligands, which is the usual scenario throughout the superfamily. A standard numbering scheme has been proposed for ZBLs [39], where metal-binding residues in motifs 2 to 5 are respectively: His/Gln116, His118, and His196 for Zn1; and Asp120, Cys221/His121, and His263 for Zn2 (*cf.* Figure 2). It is worth emphasizing that the HxHxDH motif is the hallmark of the superfamily, and such sequence diversity at motif 2 of ZBLs is rather unusual for a group of enzymes catalyzing the same reaction. This variability likely results from the strong selective pressure exerted by the comparably diverse set of β-lactam antibiotics currently in use.

Recently, new classification schemes have been proposed for ZBLs based on large-scale genomic and metagenomic data searches, suggesting that B1 and B3 ZBLs include at least five and four subgroups, respectively [40]. In addition, improved similarity criteria have been proposed for β-lactamases in general (both zinc-dependent ZBLs and serine-active enzymes), based on *ad hoc* HMM profiles [41]. The results presented here as a Pfam-based SSN and phenetic-cladistic phylogeny comparisons are consistent with those findings, stressing that B1 and B2 enzymes are more related to flavodiiron proteins (FDPs, a group of non-heme iron flavoenzymes) and alkylsulfatases, than to B3 ZBLs. FDPs like *Desulfovibrio gigas* rubredoxin:oxygen oxidoreductase ROO (Figure 1) [42] comprise a widespread family of prokaryotic oxidoreductases, containing an iron-binding MBL domain and an FMN-binding flavodoxin-like

domain [43]. ROO is a terminal reductase, which reduces O_2 to H_2O without the risk of producing reactive oxygen species. Other structurally characterized FDPs include *Moorella thermoacetica* and *Escherichia coli* nitric oxide reductases, and the *Giardia intestinalis* oxygen-scavenging enzyme. A typical His-to-Glu mutation appears at motif 2 of FDPs, located at the interface between the isoalloxazine and di-iron moieties, which likely contributes to hold the more acidic Fe(III) species. An unusual metal coordination set is found in *Thermotoga maritima* diiron oxygen sensor ODP [44], where the third histidine of motif 2 is replaced by a glutamine at motif 5. Finally, the divergent class-C type-2 FDPs from *Synechocystis* sp. display mutations at motifs 2, 3 and 4 that prevent binding of any metal ions [45].

As shown in Figure 3, alkylsulfatases belong to the same connected component as B1&B2 ZBLs. Type III sulfatases hydrolyze sulfate esters releasing HSO_4^- and the corresponding alcohol. While *Pseudomonas aeruginosa* SdsA1 [46] has preference for primary alcohol sulfates like sodium dodecylsulfate, *Pseudomonas* sp. DSM661 Pisa1 is active on secondary alcohol sulfates, which allowed the discovery that the reaction proceeds with inversion of configuration [47]. Hydrolysis of a secondary alcohol sulfate can proceed through cleavage of C–O or O–S bonds, by nucleophilic attack on the C or S atom, respectively, but only the former can result in inversion of configuration. This is an unprecedented reaction mechanism in the MBL superfamily because the nucleophilic attack occurs on the alcohol carbon by means of an S_N2 concerted reaction, where HSO_4^- is the leaving group. Thus, MBL *sec*-alkylsulfatases are highly enantioselective enzymes with great potential for application to deracemization processes [48]. In this group, there is also a clade of prokaryotic MBLs of unknown function; the human mitochondrial endoribonuclease LACTB2; and *Pseudomonas* sp. quinolone response protein PqsE. LACTB2 has been shown to use Zn(II) to hydrolyze ssRNA [49]; likely involved in RNA processing specific to mitochondrial function due to its localization and structural homology with bacterial enzymes. PqsE has been shown to bind Fe(II)/Fe(III) *in vitro* and display thiolesterase activity against a CoA-linked intermediate in the biosynthetic pathway of quinolone quorum sensing molecules, although it also contributes to the regulation of bacterial virulence through an unknown mechanism, unrelated to its thiolesterase function [50].

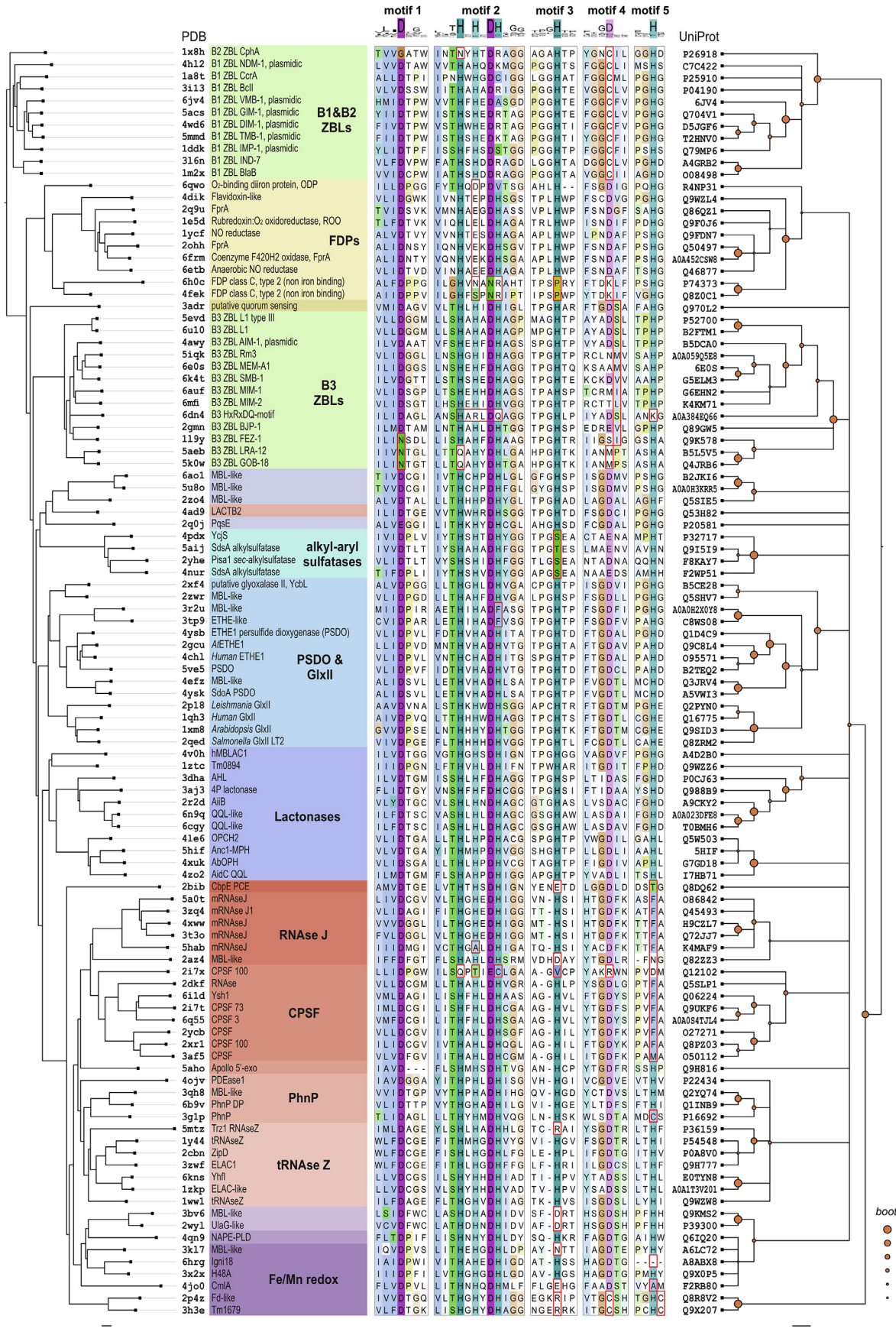
Glyoxalases II (GlxII) and persulfide dioxygenases (PSDO) share a structurally homologous MBL domain, suggestive of common ancestry. This can also be witnessed in the MSA cladogram, where this group forms a separate clade. Human glyoxalase II was the first prototypical MBL to be characterized through X-ray crystallography, disclosing the typical structural features of MBLs. GlxII are thiolesterases that convert *S*-D-lactoylglutathione into D-lactate and glutathione, as part of a ubiquitous methylglyoxal detoxification pathway [51]. The enzyme contains an αββα domain with a consensus HxHxDH motif for binding of two metal ions, reportedly Zn(II) or Mn(II), with an aspartic acid bridge in between. An additional C-terminal domain enables the enzyme to recognize and orient the glutathione moiety for proper hydrolysis, which takes place in the MBL domain metal-binding site. PSDOs are also named ETHE after the human ethylmalonic encephalopathy, a disease that has been linked to mutant PSDO enzymes [52]. Strikingly, while GlxII enzymes harbor a conventional MBL bimetallic center, PSDO enzymes have a single Fe(III) ion at site 1, even though all anticipated metal binding motif residues are conserved. Nevertheless, both enzyme groups catalyze reactions involving glutathione derivatives, *e.g.* 2-hydroxyacyl-glutathione for GlxII and glutathione-persulfide (GSS^-) for PSDOs, which detoxify sulfide by oxidation to sulfite using molecular oxygen [53]. Some PSDO enzymes like the *Burkholderia phytofirmans* enzyme are fused to rhodanese domains, working instead in sulfur assimilation pathways [54].

The next group comprises at least three phylogenetically distinct structural homologs: quorum-quenching lactonases (QQL), organophosphorus hydrolases (OPH), and human MBLAC1 endonuclease. A number of phenotypes exhibited by bacterial communities are regulated by freely diffusing small molecules signaling cell density. This quorum sensing mechanism is turned off by QQL enzymes like *Bacillus thuringensis* AiiA and *Agrobacterium* sp. AiiB, acting on *N*-acylhomoserine lactones; *Mesorhizobium japonicum* lactonase acting on 4-pyridoxolactone (an intermediate of vitamin B_6 catabolism); and *Chriseobacterium* sp. AidC lactonase. OPH enzymes like *Pseudomonas* sp. OPHC2 and methylparathion hydrolase MPH are related to QQLs but evolved to hydrolyze phosphoester bonds habitually present in organophosphorus pesticides. Indeed, OPHs may have evolved from QQLs as a resistance mechanism due to the strong selective pressure of these pesticides, resembling how ZBLs evolved to hydrolyze β-lactam antibiotics. Finally, MBLAC1 is a metazoan 3'-end mRNA processing enzyme, acting on stem-loop structures present in histone coding mRNAs [55], constituting the first of many examples of MBL nucleases.

Phosphoesterases comprise the most widespread functional group of the MBL superfamily, hydrolyzing varied phosphoesters like nucleic acids and nucleotides, phosphonates, and phospholipids. Nucleic acid processing enzymes are usually binuclear Zn(II)-dependent hydrolases, such as RNAse J, tRNAse Z, cleavage and polyadenylation specificity factors (CPSF); and DNA repair enzymes like Apollo 5'-exonuclease. These enzymes typically comprise additional domains in a modular fashion that assist the αββα hydrolytic domain at accommodating such large substrates, for instance, the tRNAse Z exosite for tRNA binding [56], β-CASP domains for binding of RNA and DNA [57] (Figure 1), and KH domains for RNA/DNA binding [58]. These modular domains can be either N-terminal, C-terminal, or inserted within the MBL fold. Indeed, the β-CASP domain sequence inserts in the loop holding the conserved His at motif 5, shifting this amino acid about 215 residues towards the C-terminus, making it difficult to find through conventional sequence alignments (*e.g. T. thermophilus* RNAse J). Analogously, the exosite insertion in tRNAse Z shifts the His at motif 5 about 75 residues to the C-terminus (*e.g. E. coli* ZipD). The yeast Trz1 tRNAse Z is an interesting example of a protein with two MBL domains where one of them evolved to improve substrate binding while losing the metal-binding and hydrolytic ability [59] (note that only the catalytic domain of Trz1 was considered in the alignment of Figure 2).

Structurally characterized phosphoesterases devoid of nuclease activity include diverse enzymes like *S. pneumoniae* modular phosphorylcholine esterase CbpE; human *N*-acyl phosphatidyl ethanolamine phospholipase D, NAPE-PLD (the only structurally characterized MBL phospholipase), and di-manganese phosphonatase PhnP from *E. coli*, part of the phosphorus scavenging CP-lyase pathway. Note that PhnP are structurally and phylogenetically related to tRNAse Z enzymes, despite their radically different functions. *Streptococcus pneumoniae* phosphorylcholinesterase CbpE is localized in the pneumococcal cell envelope [60], and catalyzes the removal the phosphorylcholine from teichoic acids, key components for cell recognition and invasiveness. The divergent *E. coli* manganese-dependent UlaG L-ascorbate-6-P lactonase clusters among phosphoesterases, and has indeed been shown to hydrolyze cyclic nucleotides [61].

Some divergent iron-dependent oxidoreductases cluster at the end of the tanglegram, including *Thermoanaerobacter tengcongensis (C. subterraneus)* Tflp, and *Streptomyces venezuelae* CmlA β-hydroxylase. Tflp contains two Cys residues in the vicinity of the di-iron center, with an Asp-to-Cys mutation at motif 4 (seen so far only in modern B1&B2 zinc-β-lactamases), plus a unique Cys residue following the His residue at motif 5. Complementary spectroscopic assays indicate that Tflp holds an [Fe–S] center under reducing conditions, and structure PDB 2p4z corresponds to an oxidized inactive form. On the other hand, CmlA is a rare β-hydroxylase clustering among phosphoesterases, which hydroxylates L-*p*-aminophenylalanine, a biosynthetic precursor of chloramphenicol.



(caption on next page)

Figure 2. Structure-function tanglegram of the MBL superfamily. Structure-guided phenogram (*left*) and the maximum-likelihood MSA-based bootstrapped consensus cladogram (*right*) of 105 selected MBLs available in the Protein Data Bank. Note that the MSA includes only conserved amino acid residues in the αββα fold, *i.e.* it does not take into account additional domains. For each dendrogram, taxa are indicated as representative PDB entries (used for structural phenogram calculation) or UniProt entries (used for MSA and cladogram calculation), respectively. A short version of the MSA is provided, comprising the corresponding sequences sorted with the tanglegram, showing the five MBL fold conserved sequence motifs as histogram logos (*top*), along with short descriptions of common protein names and families (*colored boxes*). While motif 1 contains a conserved aspartic acid residue involved in stabilization of the MBL fold near the active site; motifs 2, 3, 4 and 5 usually contain metal-coordinating residues. In general, Fe(II)/Fe(III) binding sites typical of MBL oxidoreductases exhibit more acidic residues than Zn(II) binding sites, often found in MBL hydrolases. Distinctive residues of each protein family or group are indicated in the MSA as *red* boxes. Amino acid sequence lengths are variable between these motifs, ranging 6–503 residues before motif 1 (N-terminus); 9–77 residues between motifs 1 and 2; 3–23 residues between motifs 2 and 3; 11–65 residues between motifs 3 and 4; 14–241 residues between motifs 4 and 5; and 0–58 residues after motif 5 (C-terminus). Orange dots in consensus cladogram nodes indicate bootstrap branch support values higher than 50 %.

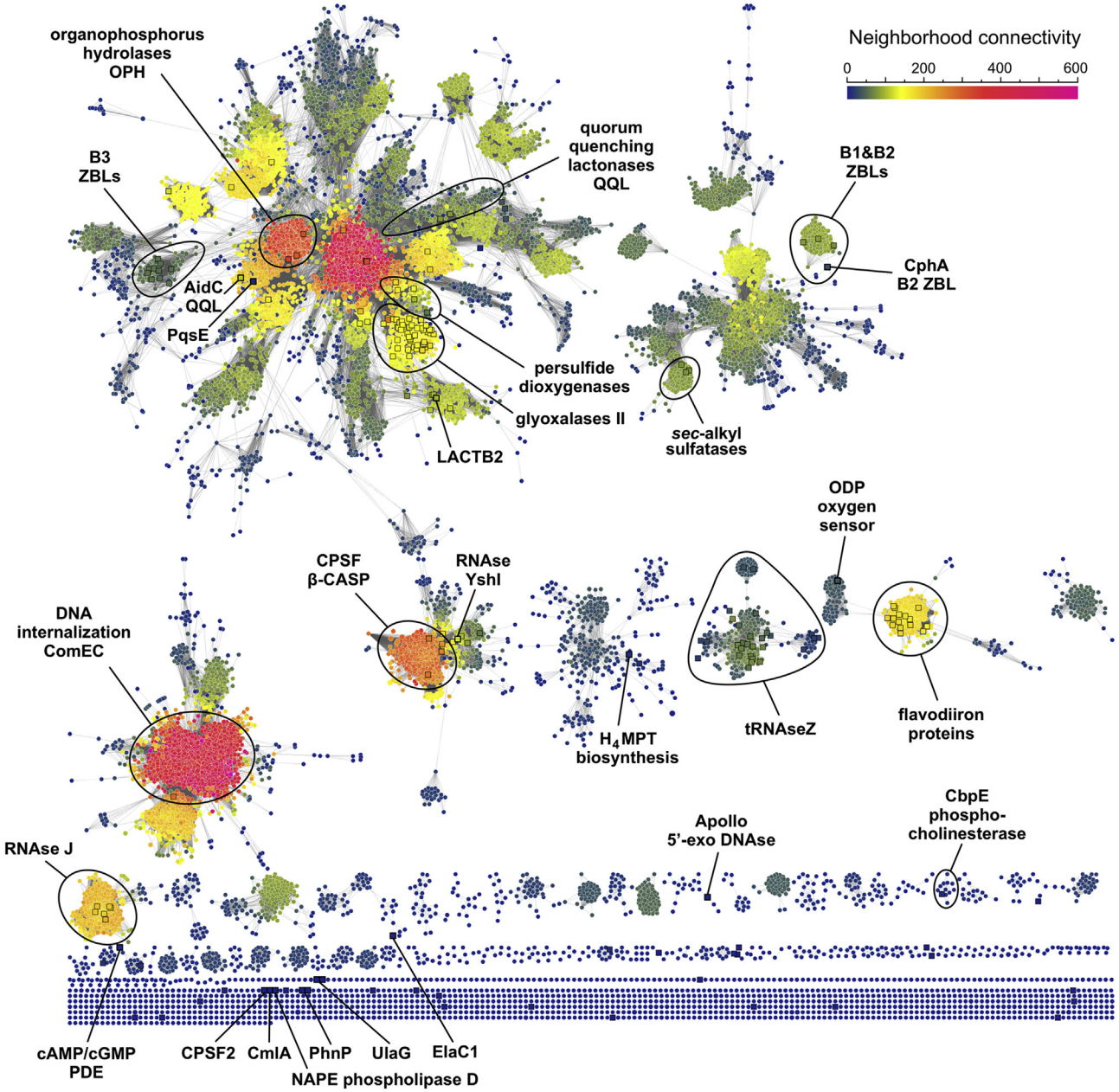


Figure 3. Sequence similarity network (SSN) for representative MBL αββα domains in the Pfam PF00753 database. The network comprises the amino acid sequence of 32,418 MBLs, expressed as 15,292 representative nodes, grouping connected nodes sharing at least 40 % sequence similarity (each representative node size is scaled by the number of proteins included). Edges between pairs of representative nodes indicate a Blast –log(*E*-value) of 20 or better, which corresponds to a sequence identity of at least ~ 30 %. Note that only MBL αββα domains were considered for Blast score calculations. For comparison, proteins and families included in the tanglegram are indicated. Square nodes indicate sequences with SwissProt and/or PDB descriptions (see Supplementary Figure). Note that many structurally and functionally characterized proteins do not cluster with the major components of the SSN but are located in isolated components (*bottom*), since their sequence similarity with proteins in major components is on average lower than 30 %. Nodes are organized with the Cytoscape Prefuse Force Directed Open CL layout, and colored by neighborhood connectivity (*top right*). See Supplementary Spreadsheets S1 and S2, and Supplementary Network for further details.

3.3. SSN analysis suggests that numerous MBL families remain to be characterized

An SSN was here calculated for the MBL superfamily using the EFI-EST webserver [33], as described in the Methods section; results are shown in Figure 3 (see Supplementary Spreadsheet S1 and Supplementary Network for full network data). SSNs are graphs with nodes representing protein sequences and edges connecting them, indicating a pairwise sequence similarity at a specified cutoff value. The metric for node similarity calculation at EFI-EST is the Blast *E*-value, which was set to –log(*E*-value) = 20. Unless otherwise stated, nodes are specifically representative nodes, which group several UniProt entries with a 40 % or higher sequence similarity, so that the SSN has fewer edges and is simpler to display graphically. By inspecting the distribution of functionally characterized proteins throughout the SSN it is evident that many MBL families remain to be characterized. In fact, one of the largest clusters in the network comprises proteins involved in DNA internalization and natural competence such as ComEC, for which no structural information is yet available and only one SwissProt entry (*Bacillus subtilis* P39695) is described. The size of connected components (CC) in the SSN follows a power law distribution, with a few clusters encompassing most nodes, and a long tail of many CCs with one or two nodes (Figure 4A). The largest CC (7259 nodes) includes glyoxalases II, PSDOs, OPHs, QQLs and B3 ZBLs; the second (1962 nodes) includes B1&B2 ZBLs and *sec*-alkyl sulfatases; and the third (1503 nodes) DNA internalization/ComEC proteins; whereas CPSF/β-CASP, tRNAse Z, RNAse J, and FDPs cluster into separate CCs of 673, 350, 333 and 297 nodes, respectively. The remaining 2915 nodes (19 %) include relatively few known MBLs sparsely scattered over 1353 smaller CCs. Analogously, the node degree shows a sharply decaying distribution, skewed towards lowly connected nodes (Figure 4B). This is probably true for all SSNs for a given alignment score cutoff, since new nodes (proteins) likely become part of existing connected components (families) instead of giving rise to new ones. Nevertheless, the curve is convex up in log-log scale (*inset*), *i.e.* it is not a power law distribution. Only 148 nodes have SwissProt descriptions and 91 nodes have at least one PDB experimentally determined structure (41 nodes have both). As depicted in Figure 3, the majority of nodes with SwissProt and PDB entries describe glyoxalases II, ribonucleases, FDPs, and ZBLs, accounting for 198 out of 15,292 nodes (1.3 %). In other words, 98.7 % of the SSN nodes need experimentally obtained functional and/or structural information so that an accurate annotation can be specified. Given the fast pace at which sequence databases grow, misannotation of macromolecular sequences is an increasingly cumbersome problem [62, 63, 64], and relying on entry annotations to define protein families is not a judicious approach.

3.4. Neighborhood connectivity distribution correlates with protein family clustering

The neighborhood connectivity (NC) statistic was introduced in 2002 by Maslov & Sneppen to describe how sets of highly connected regulatory genes control the expression of lowly connected genes [21] (Box 1). In SSNs, highly interconnected clusters share sequence and, presumably, functional similarity. Thus, members of protein families should have similar connectivities, and coloring nodes by NC provides an intuitive way of visually spotting protein families within CCs. Highly interconnected clusters indicate conserved, highly similar sequences; whereas lowly connected nodes point to rare sequences, proteins underrepresented in the SSN, or simply noise (*e.g.* truncated or incomplete sequences). For a given set of protein sequences, the SSN topology often matches the corresponding phylogenetic tree topology [12]; however, such agreement depends critically on the metrics used for network, MSA, and tree calculation [65]. This is particularly important when comparing divergent sequences sharing few conserved motifs, like the MBL superfamily. For instance, while functional families cluster into distinct clades in the tanglegram, the SSN largest connected component includes most lactonases, glyoxalases II, PSDOs, and B3 ZBLs; and separate clusters are observed for tRNAse Z, RNAse J, and CPSF phosphoesterases (Figure 3). Besides, while B1&B2 ZBLs cluster with alkylsulfatases in the SSN, the tanglegram shows that FDPs are their closest structural homologs. These

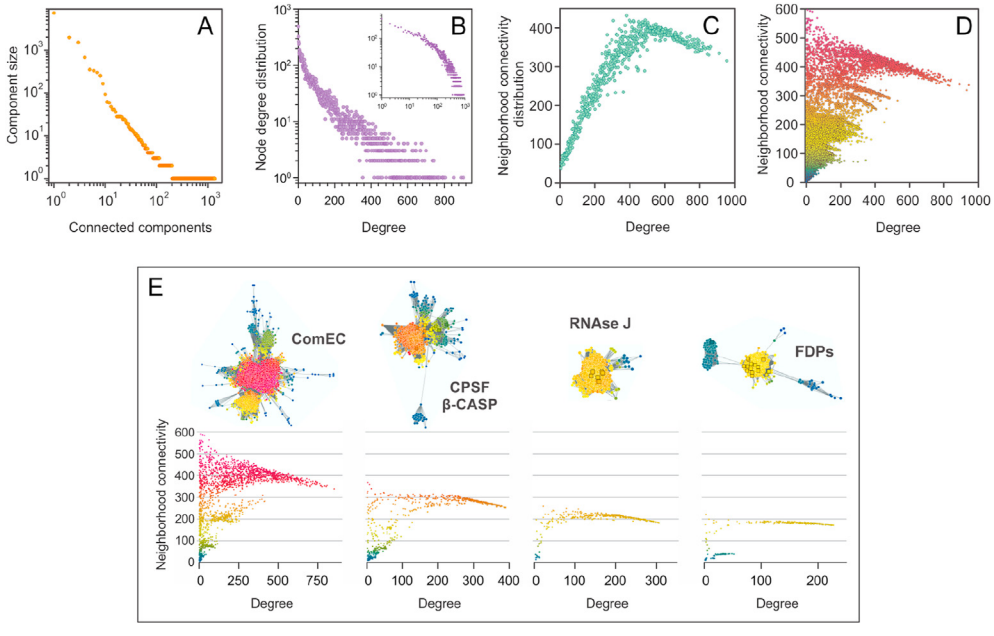
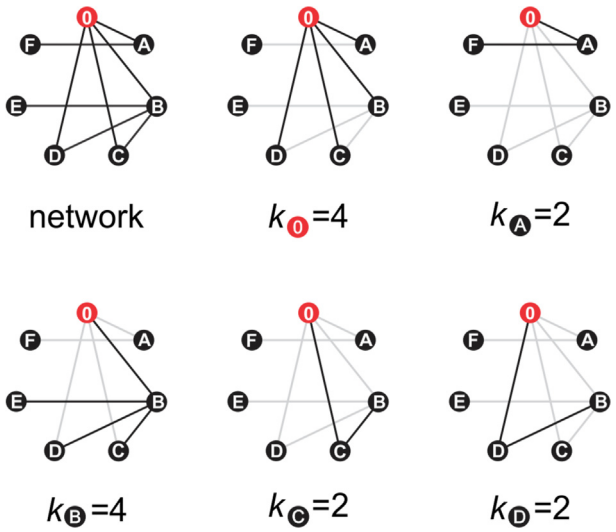


Figure 4. Topological parameters of the MBL superfamily SSN. (A) Connected components (CC) are sets of nodes connected by paths of edges. Although a full SSN comprises a single CC, setting an alignment score cutoff leads to a disconnected network aiming to isolate individual protein families, and thereby a set of CCs. The distribution of CC sizes approximately follows a power law, *i.e.* a straight line with negative slope in log-log scale. (B) Many natural networks follow a power law distribution of node degrees. However, the SSN node degree distribution is convex up and skewed toward highly connected nodes or nodes with relatively large neighborhoods (Box 1). (C) If all NC values are averaged for each degree value, the NC distribution is obtained. A maximum neighborhood connectivity of ~ 400 is observed for $k \sim 500$, which means that, on average, neighborhoods larger or smaller than ~ 500 neighbors are less interconnected. (D) Plotting all NC values for each node degree results in a scatter plot with "spikes" for highly interconnected clusters, *i.e.* highly similar groups of proteins (compare with Figure 3, the same coloring was used here). (E) Plots of NC *vs.* node degree for individual CCs provide a clearer picture of how NC values show an almost inverse linear relationship with connectivity, skewed to larger connectivity values.



Box 1. Neighborhood connectivity. Unlike many so-called "biological networks" such as protein-protein interaction networks or metabolic networks, SSNs are undirected and do not display self-edges. Then, the *neighborhood* of a node n is the set of nodes sharing an edge with n; and its *connectivity*, k_n, is the size of its neighborhood, *i.e.* the number of neighbors of n. The *degree* of node n is the number of edges reaching n, which is equivalent to k_n for SSNs. Then, the *neighborhood connectivity* (NC) of n, is defined as the average connectivity of its neighborhood, $NC_n = \Sigma(k_i)/k_n$ [21, 34]. For example, for a given node 0 (*red*) in the network {0, A, B, C, D, E, F} (*top left*), the neighborhood of 0 is {A, B, C, D} of size $k_0 = 4$, and the connectivities of each of its neighbors are $k_A = 2$, $k_B = 4$, $k_C = 2$, and $k_D = 2$. Then, the neighborhood connectivity of 0 is $NC_0 = (k_A + k_B + k_C + k_D)/k_0 = 2.5$. Note that even though nodes E and F are not neighbors of 0, they still influence its NC value by increasing k_A and k_B. Since members of a protein family are expected to cluster together sharing edges with each other, their neighborhood connectivities will exhibit comparable values. This can be readily appreciated in Figure 3 by coloring nodes according to their NC values. If N nodes in an isolated cluster are connected all-to-all, for each node $k = N - 1$ (neighbors or edges), all nodes will have a neighbor connectivity $NC = N - 1$. For example, if the network {0, A, B, C, D, E, F} had edges connecting all-to-all its $N = 7$ nodes, each node would have $k = NC = 6$ neighbors ($\sim N$ for large clusters). In other words, for highly interconnected clusters, the neighbor connectivity approaches to its maximum value, which is roughly the size of the cluster (*cf.* Figure 4 C).

apparent discrepancies likely reflect the different calculation metrics, *i.e.* Blast *E*-value for the SSN as opposed to structural homology for the tanglegram. The NC distribution reaches a maximum of ~ 400 for nodes with ~ 500 neighbors (Figure 4C), decaying almost linearly for higher connectivities. Apparently, once clusters reach a maximal connectivity or edges per node, they grow upon addition of new nodes but fewer connections are introduced. This reciprocal linear relationship observed for the full network seems to hold true also for individual clusters: plotting individual NC values reveals linear segments for each cluster, provided that enough nodes are present (Figures 4D&E). These features likely reflect the network topology arising from using the Blast *E*-value as a metric for sequence comparison, which ultimately defines the lengths of edges connecting nodes within CCs. A detailed description of these curves requires further research on SSN properties, which will shed light on the dynamics of protein network growth and degree distributions.

4. Concluding remarks

Herein, structural homology and SSN analysis are used to assist the phylogenetic reconstruction of the MBL superfamily, harnessing the protein three-dimensional arrangement of secondary structure elements as a metric for common ancestry inference. The introduced tanglegram graph disclosed structure and sequence similarity relationships between seemingly unrelated enzymes, which is suggestive of a mutual evolutionary history. Tanglegrams comprise a practical framework for protein structure-function analysis, applicable to study other protein superfamilies as well. Analogously, NC network coloring provides an intuitive picture of the distribution of protein families within the superfamily, suggesting that numerous MBL families remain to be characterized. Indeed, manually annotated entries for proteins with available experimental evidence account for only 1.3 % of the superfamily, underscoring an unfortunately frequent bias of research towards relatively few families. Automated annotation algorithms would benefit from further research on protein SSNs; establishing their topological features will give rise to improved metrics for protein function estimation.

Declarations

Author contribution statement

Javier M González: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Agency for Science and Technology Promotion (ANPCyT), grant PICT 2017-4590, Argentina.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e05867.

Acknowledgements

Dr. Liisa Holm is acknowledged for her valuable help with the Dali Lite server. J. M. G. is a staff member of Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

References

[1] A. Carfi, S. Pares, E. Duée, M. Galleni, C. Duez, J.M. Frère, O. Dideberg, The 3-D structure of a zinc metallo-beta-lactamase from Bacillus cereus reveals a new type of protein fold, EMBO J. 14 (1995) 4914–4921.
[2] N.O. Concha, B.A. Rasmussen, K. Bush, O. Herzberg, Crystal structure of the wide-spectrum binuclear zinc beta-lactamase from Bacteroides fragilis, Structure 4 (1996) 823–836.
[3] S.M. Fabiane, M.K. Sohi, T. Wan, D.J. Payne, J.H. Bateson, T. Mitchell, B.J. Sutton, Crystal structure of the zinc-dependent beta-lactamase from Bacillus cereus at 1.9 A resolution: binuclear active site with features of a mononuclear enzyme, Biochemistry 37 (1998) 12404–12411.
[4] A. Carfi, E. Duée, M. Galleni, J.M. Frère, O. Dideberg, 1.85 A resolution structure of the zinc (II) beta-lactamase from Bacillus cereus, Acta Crystallogr D Biol Crystallogr 54 (1998) 313–323.
[5] E.G. Orellano, J.E. Girardini, J.A. Cricco, E.A. Ceccarelli, A.J. Vila, Spectroscopic characterization of a binuclear metal site in Bacillus cereus beta-lactamase II. Biochemistry 37 (1998) 10173–10180.
[6] L.I. Llarrull, M.F. Tioni, J. Kowalski, B. Bennett, A.J. Vila, Evidence for a dinuclear active site in the metallo-beta-lactamase BcII with substoichiometric Co(II). A new model for metal uptake, J. Biol. Chem. 282 (2007) 30586–30595.
[7] J.M. González, M.-R. Meini, P.E. Tomatis, F.J.M. Martín, J.A. Cricco, A.J. Vila, Metallo-β-lactamases withstand low Zn(II) conditions by tuning metal-ligand interactions, Nat. Chem. Biol. 8 (2012) 698–700.
[8] L. Aravind, An evolutionary classification of the metallo-beta-lactamase fold proteins, Silico Biol. 1 (1999) 69–91.

[9] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[10] H. Daiyasu, K. Osaka, Y. Ishino, H. Toh, Expansion of the zinc metallo-hydrolase family of the beta-lactamase fold, FEBS Lett. 503 (2001) 1–6.

[11] R.D. Finn, T.K. Attwood, P.C. Babbitt, A. Bateman, P. Bork, A.J. Bridge, H.-Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G.L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D.A. Natale, M. Necci, G. Nuka, C.A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S.C. Potter, N.D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P.D. Thomas, S.C.E. Tosatto, C.H. Wu, I. Xenarios, L.-S. Yeh, S.-Y. Young, A.L. Mitchell, InterPro in 2017-beyond protein family and domain annotations, Nucleic Acids Res. 45 (2017) D190–D199.

[12] H.J. Atkinson, J.H. Morris, T.E. Ferrin, P.C. Babbitt, Using sequence similarity networks for visualization of relationships across diverse protein superfamilies, PloS One 4 (2009) e4345.

[13] H.J. Atkinson, P.C. Babbitt, An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations, PLoS Comput. Biol. 5 (2009) e1000541.

[14] F. Baier, N. Tokuriki, Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily, J. Mol. Biol. 426 (2014) 2442–2456.

[15] R. Davidson, B.-J. Baas, E. Akiva, G.L. Holliday, B.J. Polacco, J.A. LeVieux, C.R. Pullara, Y.J. Zhang, C.P. Whitman, P.C. Babbitt, A global view of structure-function relationships in the tautomerase superfamily, J. Biol. Chem. 293 (2018) 2342–2357.

[16] J.N. Copp, D.W. Anderson, E. Akiva, P.C. Babbitt, N. Tokuriki, Exploring the sequence, function, and evolutionary space of protein superfamilies using sequence similarity networks and phylogenetic reconstructions, Methods Enzymol. 620 (2019) 315–347.

[17] A. Malik, S.B. Kim, A comprehensive in silico analysis of sortase superfamily, J. Microbiol. 57 (2019) 431–443.

[18] Q. Shi, H. Wang, J. Liu, S. Li, J. Guo, H. Li, X. Jia, H. Huo, Z. Zheng, S. You, B. Qin, Old yellow enzymes: structures and structure-guided engineering for stereocomplementary bioreduction, Appl. Microbiol. Biotechnol. 104 (2020) 8155–8170.

[19] M.A. Tararina, K.N. Allen, Bioinformatic analysis of the flavin-dependent amine oxidase superfamily: adaptations for substrate specificity and catalytic diversity, J. Mol. Biol. 432 (2020) 3269–3288.

[20] D. Easley, J. Kleinberg, Networks, crowds, and markets: reasoning about a highly connected world, Cambridge University Press, 2010.

[21] S. Maslov, Specificity and stability in topology of protein networks, Science (80-) 296 (2002) 910–913.

[22] L. Holm, L.M. Laakso, Dali server update, Nucleic Acids Res. 44 (2016) W351–W355.

[23] L. Holm, Using Dali for protein structure comparison, 2020, pp. 29–42.

[24] The Uniprot Consortium & Bateman A, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.

[25] J. Pei, M. Tang, N.V. Grishin, PROMALS3D web server for accurate multiple protein sequence and structure alignments, Nucleic Acids Res. 36 (2008) W30–W34.

[26] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp, G.J. Barton, Jalview Version 2–a multiple sequence alignment editor and analysis workbench, Bioinformatics 25 (2009) 1189–1191.

[27] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.

[28] M.A. Miller, W. Pfeiffer, T. Schwartz, Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in: 2010 Gateway Computing Environments Workshop (GCE), IEEE, New Orleans, 2010, pp. 1–8.

[29] S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, Mol. Biol. Evol. 18 (2001) 691–699.

[30] C. Scornavacca, F. Zickmann, D.H. Huson, Tanglegrams for rooted phylogenetic trees and networks, Bioinformatics 27 (2011) i248–i256.

[31] D.H. Huson, C. Scornavacca, Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks, Syst. Biol. 61 (2012) 1061–1067.

[32] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682.

[33] J.A. Gerlt, J.T. Bouvier, D.B. Davidson, H.J. Imker, B. Sadkhin, D.R. Slater, K.L. Whalen, Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): a web tool for generating protein sequence similarity networks, Biochim. Biophys. Acta Protein Proteonomics 1854 (2015) 1019–1037.

[34] N.T. Doncheva, Y. Assenov, F.S. Domingues, M. Albrecht, Topological analysis and interactive visualization of biological networks and protein structures, Nat. Protoc. 7 (2012) 670–685.

[35] P. Shannon, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.

[36] G. Garau, A.M. Di Guilmi, B.G. Hall, Structure-based phylogeny of the metallo-β-lactamases, Antimicrob. Agents Chemother. 49 (2005) 2778–2784.

[37] M. Galleni, J. Lamotte-Brasseur, G.M. Rossolini, J. Spencer, O. Dideberg, J.-M. Frère, standard numbering scheme for class B β-lactamases, Antimicrob. Agents Chemother. 45 (2001) 660–663.

[38] B.G. Hall, S.J. Salipante, M. Barlow, The metallo-beta-lactamases fall into two distinct phylogenetic groups, J. Mol. Evol. 57 (2003) 249–254.

[39] G. Garau, I. García-Sáez, C. Bebrone, C. Anne, P. Mercuri, M. Galleni, J.-M. Frère, O. Dideberg, Update of the standard numbering scheme for class B β-lactamases, Antimicrob. Agents Chemother. 48 (2004) 2347–2349.

[40] F. Berglund, A. Johnning, D.G.J. Larsson, E. Kristiansson, An updated phylogeny of the metallo-β-lactamases, J. Antimicrob. Chemother. (2020).

[41] M.C. Silveira, R. Azevedo da Silva, F. Faria da Mota, M. Catanho, R. Jardim, A.C.R. Guimarães, A.B. de Miranda, Systematic identification and classification of β-lactamases based on sequence similarity criteria: β-lactamase annotation, Evol. Bioinf. Online 14 (2018), 117693431879735.

[42] C. Frazão, G. Silva, C.M. Gomes, P. Matias, R. Coelho, L. Sieker, S. Macedo, M.Y. Liu, S. Oliveira, M. Teixeira, A.V. Xavier, C. Rodrigues-Pousada, M.A. Carrondo, J. Le Gall, Structure of a dioxygen reduction enzyme from Desulfovibrio gigas, Nat. Struct. Biol. 7 (2000) 1041–1045.

[43] J.B. Vicente, M.A. Carrondo, M. Teixeira, C. Frazão, Structural studies on flavodiiron proteins, Methods Enzymol. 437 (2008) 3–19.

[44] A.R. Muok, Y. Deng, V.M. Gumerov, J.E. Chong, J.R. DeRosa, K. Kurniyati, R.E. Coleman, K.M. Lancaster, C. Li, I.B. Zhulin, B.R. Crane, A di-iron protein recruited as an Fe[II] and oxygen sensor for bacterial chemotaxis functions by stabilizing an iron-peroxy species, Proc. Natl. Acad. Sci. Unit. States Am. 116 (2019) 14955–14960.

[45] P.T. Borges, C.V. Romão, L.M. Saraiva, V.L. Gonçalves, M.A. Carrondo, M. Teixeira, C. Frazão, Analysis of a new flavodiiron core structural arrangement in Flv1-ΔFlR protein from Synechocystis sp. PCC6803, J. Struct. Biol. 205 (2019) 91–102.

[46] G. Hagelueken, T.M. Adams, L. Wiehlmann, U. Widow, H. Kolmar, B. Tummler, D.W. Heinz, W.-D. Schubert, The crystal structure of SdsA1, an alkylsulfatase from Pseudomonas aeruginosa, defines a third class of sulfatases, Proc. Natl. Acad. Sci. Unit. States Am. 103 (2006) 7631–7636.

[47] T. Knaus, M. Schober, B. Kepplinger, M. Faccinelli, J. Pitzer, K. Faber, P. Macheroux, U. Wagner, Structure and mechanism of an inverting alkylsulfatase from Pseudomonas sp. DSM6611 specific for secondary alkyl sulfates, FEBS J. 279 (2012) 4374–4384.

[48] M. Schober, P. Gadler, T. Knaus, H. Kayer, R. Birner-Grünberger, C. Gülly, P. Macheroux, U. Wagner, K. Faber, A stereoselective inverting sec -alkylsulfatase for the deracemization of sec -alcohols, Org. Lett. 13 (2011) 4296–4299.

[49] S. Levy, C.K. Allerston, V. Liveanu, M.R. Habib, O. Gileadi, G. Schuster, Identification of LACTB2, a metallo-β-lactamase protein, as a human mitochondrial endoribonuclease, Nucleic Acids Res. 44 (2016) 1813–1832.

[50] M. Zender, F. Witzgall, S.L. Drees, E. Weidel, C.K. Maurer, S. Fetzner, W. Blankenfeldt, M. Empting, R.W. Hartmann, Dissecting the multiple roles of PqsE in Pseudomonas aeruginosa virulence by discovery of small tool compounds, ACS Chem. Biol. 11 (2016) 1755–1763.

[51] A.D. Cameron, M. Ridderström, B. Olin, B. Mannervik, Crystal structure of human glyoxalase II and its complex with a glutathione thiolester substrate analogue, Structure 7 (1999) 1067–1078.

[52] V. Tiranti, C. Viscomi, T. Hildebrandt, I. Di Meo, R. Mineri, C. Tiveron, M.D. Levitt, A. Prelle, G. Fagiolari, M. Rimoldi, M. Zeviani, Loss of ETHE1, a mitochondrial dioxygenase, causes fatal sulfide toxicity in ethylmalonic encephalopathy, Nat. Med. 15 (2009) 200–205.

[53] L. Zhang, X. Liu, Z. Qin, J. Liu, Z. Zhang, Expression characteristics of sulfur dioxygenase and its function adaption to sulfide in echiuran worm Urechis unicinctus, Gene 593 (2016) 334–341.

[54] N. Motl, M.A. Skiba, O. Kabil, J.L. Smith, R. Banerjee, Structural and biochemical analyses indicate that a bacterial persulfide dioxygenase–rhodanese fusion protein functions in sulfur assimilation, J. Biol. Chem. 292 (2017) 14026–14038.

[55] I. Pettinati, P. Grzechnik, C. Ribeiro de Almeida, J. Brem, M.A. McDonough, S. Dhir, N.J. Proudfoot, C.J. Schofield, Biosynthesis of histone messenger RNA employs a specific 3' end endonuclease, Elife 7 (2018).

[56] O. Schilling, B. Späth, B. Kostelecky, A. Marchfelder, W. Meyer-Klaucke, A. Vogel, Exosite modules guide substrate recognition in the ZiPD/ElaC protein family, J. Biol. Chem. 280 (2005) 17857–17862.

[57] I. Callebaut, Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family, Nucleic Acids Res. 30 (2002) 3592–3601.

[58] N.V. Grishin, KH domain: one motif, two folds, Nucleic Acids Res. 29 (2001) 638–643.

[59] M. Ma, I. Li de la Sierra-Gallay, N. Lazar, O. Pellegrini, D. Durand, A. Marchfelder, C. Condon, H. van Tilbeurgh, The crystal structure of Trz1, the long form RNase Z from yeast, Nucleic Acids Res. 45 (2017) 6209–6216.

[60] J.A. Hermoso, L. Lagartera, A. González, M. Stelter, P. García, M. Martínez-Ripoll, J.L. García, M. Menéndez, Insights into pneumococcal pathogenesis from the crystal structure of the modular teichoic acid phosphorylcholine esterase Pce, Nat. Struct. Mol. Biol. 12 (2005) 533–538.

[61] F. Garces, F.J. Fernández, C. Montellà, E. Penya-Soler, R. Prohens, J. Aguilar, L. Baldomà, M. Coll, J. Badia, M.C. Vega, Molecular architecture of the Mn2+-dependent lactonase UlaG reveals an RNase-like metallo-β-lactamase fold and a novel quaternary structure, J. Mol. Biol. 398 (2010) 715–729.

[62] A.M. Schnoes, S.D. Brown, I. Dodevski, P.C. Babbitt, Annotation error in public databases: misannotation of molecular function in enzyme superfamilies, PLoS Comput. Biol. 5 (2009) e1000605.

[63] R. Liberal, J.W. Pinney, Simple topological properties predict functional misannotations in a metabolic network, Bioinformatics 29 (2013) i154–i161.

[64] T. Nobre, M.D. Campos, E. Lucic-Mercy, B. Arnholdt-Schmitt, Misannotation awareness: a tale of two gene-groups, Front. Plant Sci. 7 (2016).

[65] J.B. Leuthaeuser, S.T. Knutson, K. Kumar, P.C. Babbitt, J.S. Fetrow, Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site microenvironment similarity, Protein Sci. 24 (2015) 1423–1439.