Structure-guided phylogenetic reconstruction of the superfamily of proteins sharing the metallo-β-lactamase fold

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.

Email: javierg@conicet.gov.ar

http://orcid.org/0000-0002-3298-2235

Keywords: metallo-lactamase, protein superfamily, tanglegram, sequence similarity network, neighborhood connectivity

Abbreviations: MBL (metallo-β-lactamase), MSA (multiple sequence alignment), SSN (sequence similarity network), PSDO (persulfide dioxygenase), FDP (flavo diiron protein), QQL (quorum quenching lactonase), OPH (organophosphorus hydrolase), ZBL (zinc-β-lactamase), CPSF (cleavage and polyadenylation specificity factor), CC (connected component), NC (neighborhood connectivity).

ABSTRACT

The superfamily of metallo-β-lactamases (MBL) comprises an ancient group of proteins found in all domains of life, sharing a characteristic αββα fold and a histidine-rich motif for binding of transition metal ions, with the ability to catalyze a variety of hydrolysis and redox reactions. Herein, structural homology and sequence similarity network (SSN) analysis are used to assist the phylogenetic reconstruction of the MBL superfamily, introducing tanglegrams to evaluate structure-function relationships. SSN neighborhood connectivity is applied for spotting protein families within SSN clusters, showing that 98 % of the superfamily remains to be explored experimentally. Further SSN research is suggested in order to determine their topological properties, which will be instrumental for the improvement of automated sequence annotation methods.

The metallo-β-lactamase (MBL) superfamily comprises an ancient group of proteins found in all domains of life, sharing a characteristic αββα fold and a histidine-rich motif for binding of transition metal ions. 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 [3-5], 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]. 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 will be followed here to define any protein with at least one characteristic MBL domain, leaving the acronym ZBL to describe metallo-β-lactamases themselves.

A great diversity of proteins evolved in the MBL superfamily by combining catalytic MBL domains and substrate recognition domains in a modular fashion. Besides, 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 contain αββα domains through X-ray crystallography, whereas the InterPro 77.0 database [11] 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 this report, the Pfam and Protein Data Bank databases are interrogated in order to obtain an updated picture of the distribution of proteins throughout the MBL superfamily and to get insights into their structure-function relationships. It is shown that the most widespread MBLs are those acting on phosphoesters, such as nucleic acids, nucleotides, phospholipids and phosphonates; whereas the more specialized enzymes are less distributed, such as zinc-β-lactamases, flavo-diiron oxidoreductases, and MBLs involved in secondary metabolism, with the important exception of the ubiquitous glyoxalases II. In addition, the structural diversity of membrane-associated MBL proteins has not been explored by experimental structural approaches, despite being involved in fundamental processes like natural transformation and horizontal gene transfer.

MATERIALS AND METHODS

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 [12], 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 [12]. Next, the full amino acid sequence corresponding to each of these 105 structures were retrieved from the

UniProt database [13], 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 [14]. This MSA was manually edited with Jalview 2.9 [15] 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 [16], running at the Cipres server [17]. A best-scoring bootstrapped tree was obtained after 1002 replicates, using the WAG substitution matrix as evolutionary model [18], 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 [19], available in Dendroscope 3.5.9 [20], 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).

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) [21]. The RP55 representative proteome MSA (62,213 sequences by 1,251 columns) was downloaded and manually edited with Jalview 2.9 [15], 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 [13] and reduced to a final set of 32,418 sequences, by applying a 70 % similarity cutoff with CD-Hit [22] and ensuring that all 105 sequences present in the tanglegram were included. A sequence similarity network (SSN) [23] was then calculated with this 32,418-sequence dataset, using the EFI-EST online tool [24]. 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 [25], as implemented in Cytoscape 3.7.1 [26]. Network statistics plots were prepared with SigmaPlot 12 (Systat Software). All figures were prepared with Corel Draw X7 (Corel).

RESULTS AND DISCUSSION

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.

Here, in order to assist the phylogenetic reconstruction of the MBL superfamily, two

trees were calculated and compared for a selected set of MBLs with available experimentally determined structures. First, 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. 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).

Phenetic and cladistic considerations shed light on mutual MBL ancestors

ZBLs comprise a divergent polyphyletic group of MBLs, comprising subclasses B1, B2, and B3 [27]. As shown in the tanglegram and suggested previously [28], 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. 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. 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.

Notably, the closest structural homologs of B1&B2 ZBLs are not B3 ZBLs but a group of non-heme iron flavoenzymes, known as flavodiiron proteins (FDPs). FDPs like *Desulfovibrio gigas* rubredoxin:oxygen oxidoreductase ROO (Figure 1) [29] comprise a widespread family of prokaryotic oxidoreductases, containing an iron-binding MBL domain and an FMN-binding flavodoxin-like domain [30]. 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 [31], 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 [32].

The next group comprises phylogenetically distinct proteins structurally related to alkylsulfatases. Type III sulfatases hydrolyze sulfate esters releasing HSO_4^- and the corresponding alcohol. While *Pseudomonas aeruginosa* SdsA1 [33] 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 [34]. 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 [35]. 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 [36]; 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 [37].

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 [38]. 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 [39]. 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 [40]. Some PSDO enzymes like the *Burkholderia phytofirmans* enzyme are fused to rhodanese domains, working instead in sulfur assimilation pathways [41].

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 [42], 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 [43], β-CASP domains for ssRNA binding [44] (Figure 1), and KH domains for RNA/DNA binding [45]. 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 [46] (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 phosphoryl-cholinesterase CbpE is localized in the pneumococcal cell envelope [47], 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 [48].

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.

SSN analysis suggests that numerous MBL families remain to be characterized

Sequence similarity networks (SSN) have been applied over the past decade to analyze a number of protein superfamilies [23, 49]. Herein, an SSN was calculated for the MBL superfamily using the EFI-EST webserver [24], as described in the Methods section; results are shown in Figure 3. 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). 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 4 A). 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 4 B). 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 [50-52], and relying on entry annotations to define protein families is not a judicious approach.

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 [53] (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 [23]; however, such agreement depends critically on the metrics used for network, MSA, and tree calculation [54]. 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 apparent discrepancies likely reflect the different calculation metrics, *i.e.* sequence similarity 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 4 C), 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 4 D&E).

Concluding remarks

Herein, structural homology and sequence similarity network 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. In addition, such tanglegrams provide a practical framework for protein structure-function analysis, applicable to study other protein superfamilies as well. Neighborhood connectivity distribution considerations indicate that numerous MBL families remain to be characterized. Besides, 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 MBLs. Indeed, automated annotation algorithms would benefit from further research on protein sequence similarity networks, since establishing their topological features will give rise to improved metrics for protein function estimation.

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).

References

1. Carfi, A., Pares, S., Duee, E., Galleni, M., Duez, C., Frere, J. M. & Dideberg, O. (1995) The 3-D structure of a zinc metallo-beta-lactamase from *Bacillus cereus* reveals a new type of protein fold, *The EMBO Journal.* **14**, 4914-21.
2. Concha, N. O., Rasmussen, B. A., Bush, K. & Herzberg, O. (1996) Crystal structure of the wide-spectrum binuclear zinc beta-lactamase from Bacteroides fragilis, *Structure.* **4**, 823-36.
3. Fabiane, S. M., Sohi, M. K., Wan, T., Payne, D. J., Bateson, J. H., Mitchell, T. & Sutton, B. J. (1998) 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**, 12404-11.
4. Carfi, A., Duee, E., Galleni, M., Frere, J. M. & Dideberg, O. (1998) 1.85 A resolution structure of the zinc (II) beta-lactamase from *Bacillus cereus*, *Acta Crystallographica Section D Biological Crystallography.* **54**, 313-23.
5. Orellano, E. G., Girardini, J. E., Cricco, J. A., Ceccarelli, E. A. & Vila, A. J. (1998) Spectroscopic characterization of a binuclear metal site in *Bacillus cereus* beta-lactamase II, *Biochemistry.* **37**, 10173-80.
6. Llarrull, L. I., Tioni, M. F., Kowalski, J., Bennett, B. & Vila, A. J. (2007) Evidence for a dinuclear active site in the metallo-beta-lactamase BcII with substoichiometric Co(II). A new model for metal uptake, *The Journal Biological Chemistry.* **282**, 30586-95.
7. González, J. M., Meini, M. R., Tomatis, P. E., Medrano Martin, F. J., Cricco, J. A. & Vila, A. J. (2012) Metallo-β-lactamases withstand low Zn(II) conditions by tuning metal-ligand interactions, *Nature Chemical Biology.* **8**, 698-700.
8. Aravind, L. (1999) An evolutionary classification of the metallo-beta-lactamase fold proteins, *In Silico Biol.* **1**, 69-91.
9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research.* **25**, 3389-402.
10. Daiyasu, H., Osaka, K., Ishino, Y. & Toh, H. (2001) Expansion of the zinc metallo-hydrolase family of the beta-lactamase fold, *FEBS Letters.* **503**, 1-6.
11. Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, Silvio C E., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y. & Mitchell, A. L. (2017) InterPro in 2017-beyond protein family and domain annotations, *Nucleic Acids Research.* **45**, D190-D199.
12. Holm, L. & Laakso, L. M. (2016) Dali server update, *Nucleic Acids Research.* **44**, W351-5.
13. Jungo, F., Bougueleret, L., Xenarios, I. & Poux, S. (2012) The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data, *Toxicon.* **60**, 551-557.
14. Pei, J., Kim, B. H. & Grishin, N. V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments, *Nucleic Acids Research.* **36**, 2295-300.
15. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench, *Bioinformatics.* **25**, 1189-91.
16. Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics.* **22**, 2688-90.
17. Miller, M. A., Pfeiffer, W. & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Paper presented at the *2010 Gateway Computing Environments Workshop (GCE)*.
18. Whelan, S. & Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Molecular Biology and Evolution.* **18**, 691-9.

19. Scornavacca, C., Zickmann, F. & Huson, D. H. (2011) Tanglegrams for rooted phylogenetic trees and networks, *Bioinformatics.* **27**, i248-56.
20. Huson, D. H. & Scornavacca, C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks, *Systematic Biology.* **61**, 1061-7.
21. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C E. & Finn, R. D. (2018) The Pfam protein families database in 2019, *Nucleic Acids Research.* **47**, D427-D432.
22. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics.* **26**, 680-682.
23. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies, *PLoS One.* **4**, e4345.
24. Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R. & Whalen, K. L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks, *Biochimica et Biophysica Acta.* **1854**, 1019-37.
25. Doncheva, N. T., Assenov, Y., Domingues, F. S. & Albrecht, M. (2012) Topological analysis and interactive visualization of biological networks and protein structures, *Nature Protocols.* **7**, 670-85.
26. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.* **13**, 2498-504.
27. Galleni, M., Lamotte-Brasseur, J., Rossolini, G. M., Spencer, J., Dideberg, O. & Frère, J. M. (2001) Standard numbering scheme for class B beta-lactamases, *Antimicrobial Agents and Chemotherapy.* **45**, 660-3.
28. Hall, B. G., Salipante, S. J. & Barlow, M. (2003) The metallo-beta-lactamases fall into two distinct phylogenetic groups, *Journal of Molecular Evolution.* **57**, 249-54.
29. Frazao, C., Silva, G., Gomes, C. M., Matias, P., Coelho, R., Sieker, L., Macedo, S., Liu, M. Y., Oliveira, S., Teixeira, M., Xavier, A. V., Rodrigues-Pousada, C., Carrondo, M. A. & Le Gall, J. (2000) Structure of a dioxygen reduction enzyme from *Desulfovibrio gigas*, *Nature Structural Biology.* **7**, 1041-5.
30. Vicente, J. B., Carrondo, M. A., Teixeira, M. & Frazao, C. (2008) Structural studies on flavodiiron proteins, *Methods in Enzymology.* **437**, 3-19.
31. Muok, A. R., Deng, Y., Gumerov, V. M., Chong, J. E., DeRosa, J. R., Kurniyati, K., Coleman, R. E., Lancaster, K. M., Li, C., Zhulin, I. B. & Crane, B. R. (2019) A di-iron protein recruited as an Fe[II] and oxygen sensor for bacterial chemotaxis functions by stabilizing an iron-peroxy species, *Proceedings of the National Academy of Sciences U S A.* **116**, 14955-14960.
32. Borges, P. T., Romao, C. V., Saraiva, L. M., Goncalves, V. L., Carrondo, M. A., Teixeira, M. & Frazao, C. (2019) Analysis of a new flavodiiron core structural arrangement in Flv1-DeltaFlR protein from *Synechocystis* sp. PCC6803, *Journal of Structural Biology.* **205**, 91-102.
33. Hagelueken, G., Adams, T. M., Wiehlmann, L., Widow, U., Kolmar, H., Tummler, B., Heinz, D. W. & Schubert, W. D. (2006) The crystal structure of SdsA1, an alkylsulfatase from *Pseudomonas aeruginosa*, defines a third class of sulfatases, *Proceedings of the National Academy of Sciences U S A.* **103**, 7631-6.
34. Knaus, T., Schober, M., Kepplinger, B., Faccinelli, M., Pitzer, J., Faber, K., Macheroux, P. & Wagner, U. (2012) Structure and mechanism of an inverting alkylsulfatase from *Pseudomonas* sp. DSM6611 specific for secondary alkyl sulfates, *The FEBS Journal.* **279**, 4374-84.
35. Schober, M., Gadler, P., Knaus, T., Kayer, H., Birner-Grünberger, R., Gülly, C., Macheroux, P., Wagner, U. & Faber, K. (2011) A Stereoselective Inverting sec-Alkylsulfatase for the Deracemization of sec-Alcohols, *Organic Letters.* **13**, 4296-4299.
36. Levy, S., Allerston, C. K., Liveanu, V., Habib, M. R., Gileadi, O. & Schuster, G. (2016) Identification of LACTB2, a metallo-β-lactamase protein, as a human mitochondrial endoribonuclease, *Nucleic Acids Research.* **44**, 1813-1832.
37. Zender, M., Witzgall, F., Drees, S. L., Weidel, E., Maurer, C. K., Fetzner, S., Blankenfeldt, W., Empting, M. & Hartmann, R. W. (2016) Dissecting the Multiple Roles of PqsE in Pseudomonas aeruginosa Virulence by Discovery of Small Tool Compounds, *ACS Chemical Biology.* **11**, 1755-1763.
38. Cameron, A. D., Ridderstrom, M., Olin, B. & Mannervik, B. (1999) Crystal structure of human

glyoxalase II and its complex with a glutathione thiolester substrate analogue, *Structure.* **7**, 1067-78.
39. Tiranti, V., Viscomi, C., Hildebrandt, T., Di Meo, I., Mineri, R., Tiveron, C., Levitt, M. D., Prelle, A., Fagiolari, G., Rimoldi, M. & Zeviani, M. (2009) Loss of ETHE1, a mitochondrial dioxygenase, causes fatal sulfide toxicity in ethylmalonic encephalopathy, *Nature Medicine.* **15**, 200-5.
40. Zhang, L., Liu, X., Qin, Z., Liu, J. & Zhang, Z. (2016) Expression characteristics of sulfur dioxygenase and its function adaption to sulfide in echiuran worm *Urechis unicinctus*, *Gene.* **593**, 334-41.
41. Motl, N., Skiba, M. A., Kabil, O., Smith, J. L. & Banerjee, R. (2017) Structural and biochemical analyses indicate that a bacterial persulfide dioxygenase-rhodanese fusion protein functions in sulfur assimilation, *The Journal of Biological Chemistry.* **292**, 14026-14038.
42. Pettinati, I., Grzechnik, P., Ribeiro de Almeida, C., Brem, J., McDonough, M. A., Dhir, S., Proudfoot, N. J. & Schofield, C. J. (2018) Biosynthesis of histone messenger RNA employs a specific 3' end endonuclease, *eLife.* **7**, e39865.
43. Schilling, O., Spath, B., Kostelecky, B., Marchfelder, A., Meyer-Klaucke, W. & Vogel, A. (2005) Exosite modules guide substrate recognition in the ZiPD/ElaC protein family, *The Journal of Biological Chemistry.* **280**, 17857-62.
44. Callebaut, I., Moshous, D., Mornon, J. P. & de Villartay, J. P. (2002) Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family, *Nucleic Acids Res.* **30**, 3592-601.
45. Grishin, N. V. (2001) KH domain: one motif, two folds, *Nucleic Acids Research.* **29**, 638-43.
46. Xie, R., Zhang, H., Wang, X. Z., Yang, X. Z., Wu, S. N., Wang, H. G., Shen, P. & Ma, T. H. (2017) The protective effect of betulinic acid (BA) diabetic nephropathy on streptozotocin (STZ)-induced diabetic rats, *Food & Function.* **8**, 299-306.
47. Hermoso, J. A., Lagartera, L., Gonzalez, A., Stelter, M., Garcia, P., Martinez-Ripoll, M., Garcia, J. L. & Menendez, M. (2005) Insights into pneumococcal pathogenesis from the crystal structure of the modular teichoic acid phosphorylcholine esterase Pce, *Nature Structural and Molecular Biology.* **12**, 533-8.
48. Garces, F., Fernandez, F. J., Montella, C., Penya-Soler, E., Prohens, R., Aguilar, J., Baldoma, L., Coll, M., Badia, J. & Vega, M. C. (2010) Molecular architecture of the Mn^{2+}-dependent lactonase UlaG reveals an RNase-like metallo-beta-lactamase fold and a novel quaternary structure, *Journal of Molecular Biology.* **398**, 715-29.
49. Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L. & Babbitt, P. C. (2014) The Structure–Function Linkage Database, *Nucleic Acids Research.* **42**, D521-D530.
50. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. (2009) Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies, *PLoS Computational Biology.* **5**, e1000605.
51. Liberal, R. & Pinney, J. W. (2013) Simple topological properties predict functional misannotations in a metabolic network, *Bioinformatics.* **29**, i154-i161.
52. Nobre, T., Campos, M. D., Lucic-Mercy, E. & Arnholdt-Schmitt, B. (2016) Misannotation Awareness: A Tale of Two Gene-Groups, *Frontiers in Plant Science.* **7**, 868.
53. Maslov, S. & Sneppen, K. (2002) Specificity and stability in topology of protein networks, *Science.* **296**, 910-3.
54. Leuthaeuser, J. B., Knutson, S. T., Kumar, K., Babbitt, P. C. & Fetrow, J. S. (2015) Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site microenvironment similarity, *Protein Science.* **24**, 1423-39.

Figures

Figure 1. **Structural diversity of MBLs**. ZBLs like plasmid-borne *Klebsiella pneumoniae* NDM-1 (PDB 4hl2, *left*) comprise a single αββα domain, 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 for single-stranded RNA binding. Finally, MBL oxidoreductases such as the flavo-diiron protein ROO from *Desulfovibrio gigas* (PDB 1e5d, *right*) utilize non-heme Fe(II)/Fe(III) for catalysis, exhibiting a more acidic metal ligand set, in combination with an FMN-binding flavodoxin domain, displaying a homodimeric quaternary structure. Metal ions are indicated as numbered spheres. Amino acid side chains follow the coloring scheme of Figure 2.

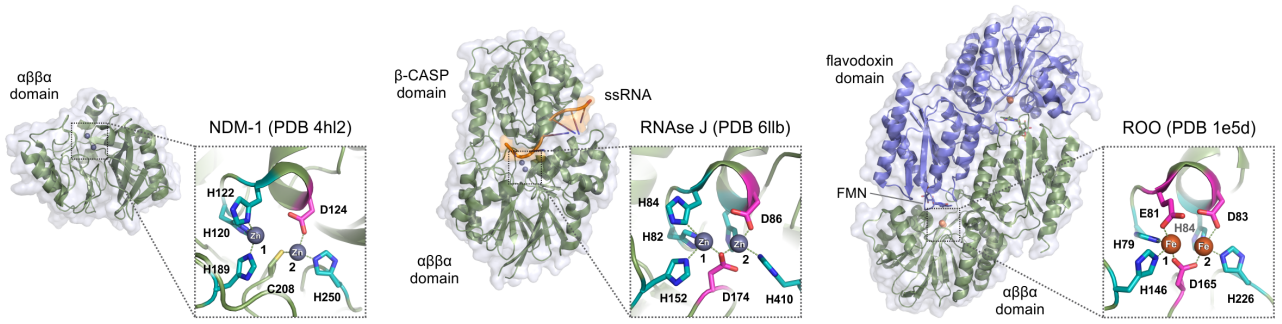


Figure 2. **Structure-function tanglegram of the MBL superfamily**. Structure-guided phenogram (*left*) and the maximum-likelihood MSA-based bootstrapped consensus cladogram (*right*) of 105 selected MBLs available in the Protein Data Bank. Note that the MSA includes only conserved amino acid residues in the αββα fold, *i.e.* 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. Unusual residues in the MSA are indicated as *red* boxes. Amino acid sequence lengths are variable between these motifs, ranging 6-503 before motif 1 (N-terminus); 9-77 between motifs 1 and 2; 3-23 between motifs 2 and 3; 11-65 between motifs 3 and 4; 14-241 between motifs 4 and 5; and 0-58 after motif 5 (C-terminus). Orange dots in cladogram nodes indicate bootstrap branch support values higher than 50 %.

(*Next page*).

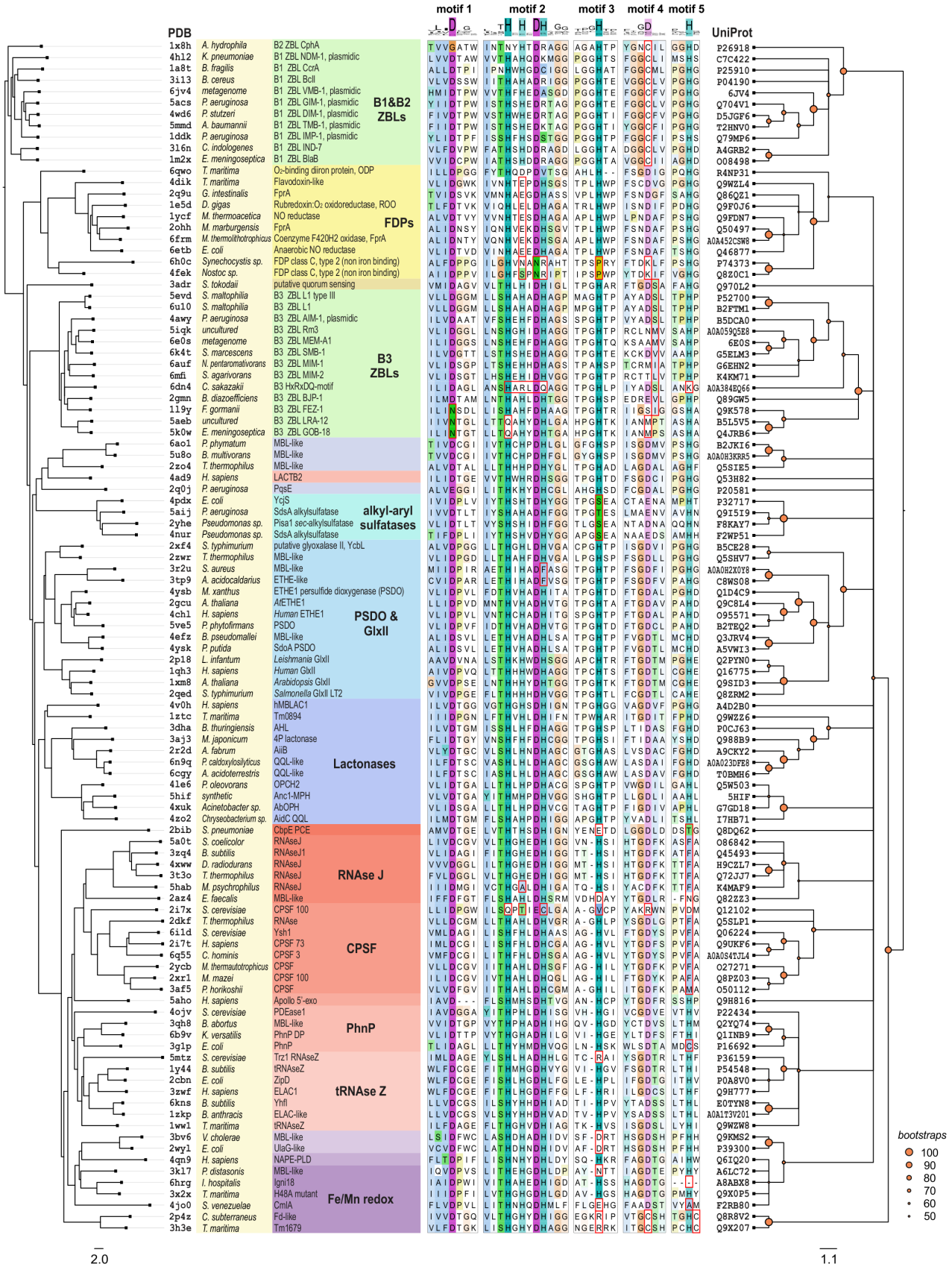
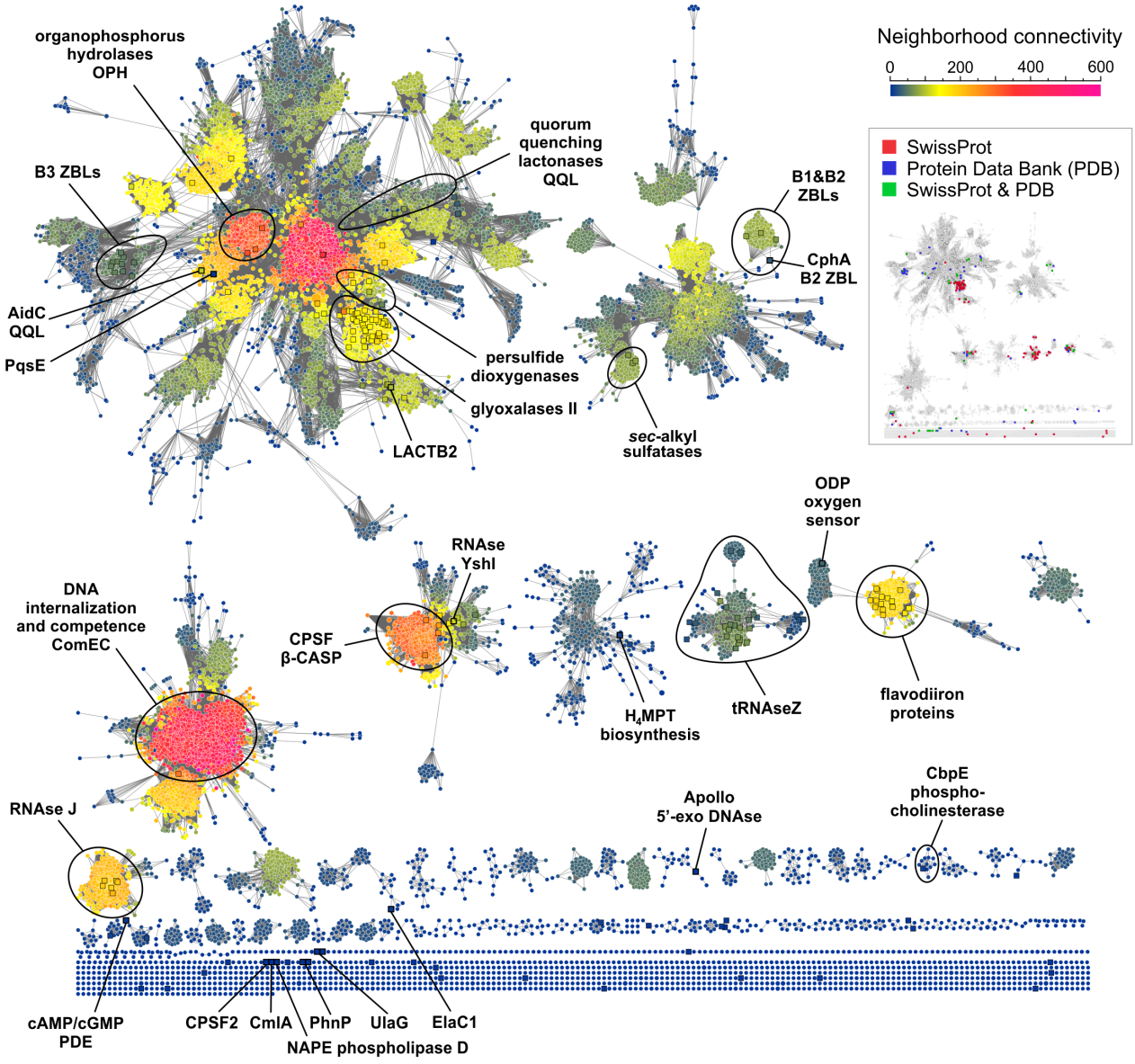

motif 1
motif 2
motif 3
motif 4
motif 5
PDB
UniProt
1x8h A. hydrophila B2 ZBL CphA P26918
4hl2 K. pneumoniae B1 ZBL NDM-1, plasmidic C7C422
1a8t B. fragilis B1 ZBL CcrA P25910
3i13 B. cereus B1 ZBL BcII P04190
6jv4 metagenome B1 ZBL VMB-1, plasmidic 6JV4
5acs P. aeruginosa B1 ZBL GIM-1, plasmidic Q704V1
4wd6 P. stutzeri B1 ZBL DIM-1, plasmidic D5JGF6
5mmd A. baumannii B1 ZBL TMB-1, plasmidic T2HNV0
1ddk P. aeruginosa B1 ZBL IMP-1, plasmidic Q79MP6
3l6n C. indologenes B1 ZBL IND-7 A4GRB2
1m2x E. meningoseptica B1 ZBL BlaB O08498
B1&B2 ZBLs
6qwo T. maritima O2-binding diiron protein, ODP R4NP31
4dik T. maritima Flavodoxin-like Q9WZL4
2q9u G. intestinalis FprA Q86QZ1
1e5d D. gigas Rubredoxin:O2 oxidoreductase, ROO Q9F0J6
1ycf M. thermoacetica NO reductase Q9FDN7
2ohh M. marburgensis FprA Q50497
6frm M. thermolithotrophicus Coenzyme F420H2 oxidase, FprA A0A452CSW8
6etb E. coli Anaerobic NO reductase Q46877
6h0c Synechocystis sp. FDP class C, type 2 (non iron binding) P74373
4fek Nostoc sp. FDP class C, type 2 (non iron binding) Q8ZOC1
FDPs
3adr S. tokodaii putative quorum sensing Q970L2
5evd S. maltophilia B3 ZBL L1 type III P52700
6u10 S. maltophilia B3 ZBL L1 B2FTM1
4awy P. aeruginosa B3 ZBL AIM-1, plasmidic B5DCA0
5iqk uncultured B3 ZBL Rm3 A0A059Q5E8
6e0s metagenome B3 ZBL MEM-A1 6E0S
6k4t S. marcescens B3 ZBL SMB-1 G5ELM3
6auf N. pentaromativorans B3 ZBL MIM-1 G6EHN2
6mfi S. agarivorans B3 ZBL MIM-2 K4KM71
6dn4 C. sakazakii B3 HxRxDQ-motif A0A384EQ66
2gmn B. diazoefficiens B3 ZBL BJP-1 Q89GW5
1l9y F. gormanii B3 ZBL FEZ-1 Q9K578
5aeb uncultured B3 ZBL LRA-12 B5L5V5
5k0w E. meningoseptica B3 ZBL GOB-18 Q4JRB6
B3 ZBLs
6aol P. phymatum MBL-like B2JKI6
5u8o B. multivorans MBL-like A0A0H3KRR5
2zo4 T. thermophilus MBL-like Q5SIE5
4ad9 H. sapiens LACTB2 Q53H82
2q0j P. aeruginosa PqsE P20581
4pdx E. coli YcjS P32717
5aij P. aeruginosa SdsA alkylsulfatase Q9I5I9
2yhe Pseudomonas sp. Pisa1 sec-alkylsulfatase F8KAY7
4nur Pseudomonas sp. SdsA alkylsulfatase F2WP51
alkyl-aryl sulfatases
2xf4 S. typhimurium putative glyoxalase II, YcbL B5CE28
2zwr T. thermophilus MBL-like Q5SHV7
3r2u S. aureus MBL-like A0A0H2X0Y8
3tp9 A. acidocaldarius ETHE-like C8WS08
4ysb M. xanthus ETHE1 persulfide dioxygenase (PSDO) Q1D4C9
2gcu A. thaliana AtETHE1 Q9C8L4
4chl H. sapiens Human ETHE1 O95571
5ve5 P. phytofirmans PSDO B2TEQ2
4efz B. pseudomallei MBL-like Q3JRV4
4ysk P. putida SdoA PSDO A5VWI3
2p18 L. infantum Leishmania GlxII Q2PYN0
1qh3 H. sapiens Human GlxII Q16775
1xm8 A. thaliana Arabidopsis GlxII Q9SID3
2qed S. typhimurium Salmonella GlxII LT2 Q8ZRM2
PSDO & GlxII
4v0h H. sapiens hMBLAC1 A4D2B0
1ztc T. maritima Tm0894 Q9WZZ6
3dha B. thuringiensis AHL P0CJ63
3aj3 M. japonicum 4P lactonase Q988B9
2r2d A. fabrum AiiB A9CKY2
6n9q P. caldoxylosilyticus QQL-like A0A023DFE8
6cgy A. acidoterrestris QQL-like T0BMH6
4le6 P. oleovorans OPCH2 Q5W503
5hif synthetic Anc1-MPH 5HIF
4xuk Acinetobacter sp. AbOPH G7GD18
4zo2 Chryseobacterium sp. AidC QQL I7HB71
Lactonases
2bib S. pneumoniae CbpE PCE Q8DQ62
5a0t S. coelicolor RNAseJ O86842
3zq4 B. subtilis RNAseJ1 Q45493
4xww D. radiodurans RNAseJ H9CZL7
3t3o T. thermophilus RNAseJ Q72JJ7
5hab M. psychrophilus RNAseJ K4MAF9
2az4 E. faecalis MBL-like Q82ZZ3
RNAse J
2i7x S. cerevisiae CPSF 100 Q12102
2dkf T. thermophilus RNAse Q5SLP1
6i1d S. cerevisiae Ysh1 Q06224
2i7t H. sapiens CPSF 73 Q9UKF6
6q55 C. hominis CPSF 3 A0A0S4TJL4
2ycb M. thermautotrophicus CPSF O27271
2xr1 M. mazei CPSF 100 Q8PZ03
3af5 P. horikoshii CPSF O50112
5aho H. sapiens Apollo 5'-exo Q9H816
CPSF
4ojv S. cerevisiae PDEase1 P22434
3qh8 B. abortus MBL-like Q2YQ74
6b9v K. versatilis PhnP DP Q1INB9
3g1p E. coli PhnP P16692
PhnP
5mtz S. cerevisiae Trz1 RNAseZ P36159
1y44 B. subtilis tRNAseZ P54548
2cbn E. coli ZipD P0A8V0
2zwf H. sapiens ELAC1 Q9H777
6kns B. subtilis Yhfl E0TYN8
1zkp B. anthracis ELAC-like A0A1T3V201
1ww1 T. maritima tRNAseZ Q9WZW8
3bv6 V. cholerae MBL-like Q9KMS2
2wyl E. coli UlaG-like P39300
4qn9 H. sapiens NAPE-PLD Q6IQ20
3kl7 P. distasonis MBL-like A6LC72
6hrg I. hospitalis Igni18 A8ABX8
3x2x T. maritima H48A mutant Q9X0P5
4jo0 S. venezuelae CmlA F2RB80
2p4z C. subterraneus Fd-like Q8R8V2
3h3e T. maritima Tm1679 Q9X207
tRNAse Z
Fe/Mn redox
bootstraps
100
90
80
70
60
50
2.0
1.1

Figure 3. Sequence similarity network (SSN) for representative MBL αββα domains in the Pfam 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 (*inset*). 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*).



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, $NC_n = \Sigma(k_i)/k_n$ [25, 53]. 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), and 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 4C).

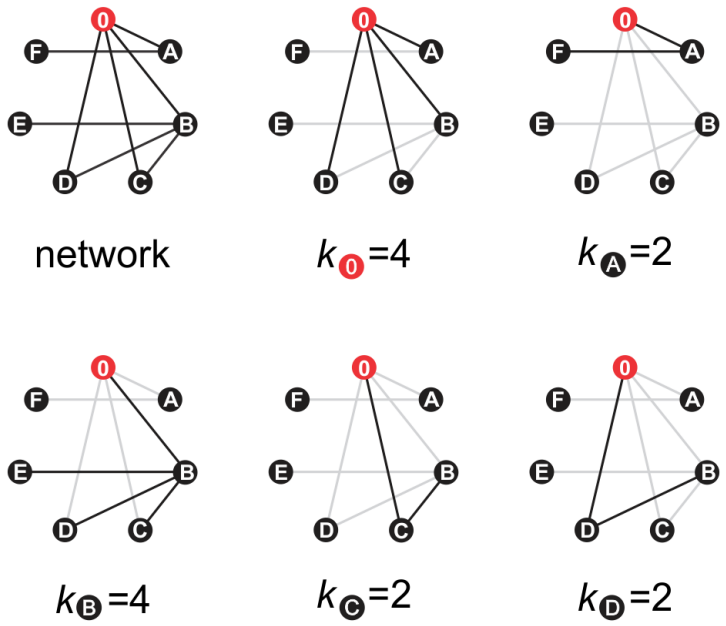
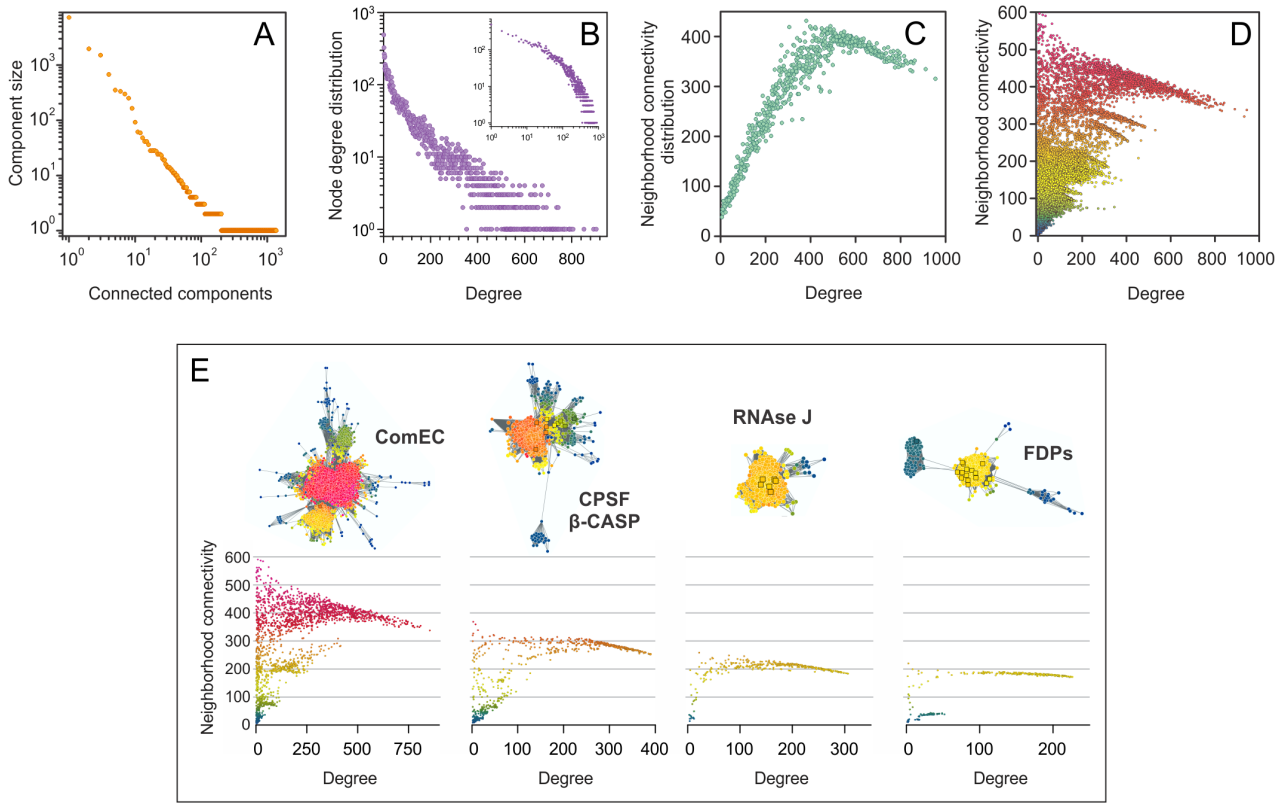



Figure 4. **Topological parameters of the MBL superfamily SSN**. (**A**) Connected components (CC) are sets of nodes connected by paths of edges. Although a full SSN comprises a single CC, setting an alignment score cutoff leads to a disconnected network aiming to isolate individual protein families, and thereby a set of CCs. The distribution of CC sizes 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 individual 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.