Discovery of New Enzymatic Functions and Metabolic Pathways Using Genomic Enzymology Web Tools

Remi Zallot[1,2], Nils C. Oberg[1], and John A. Gerlt[1,3*]

[1]Carl. R. Woese Institute for Genomic Biology, University of Illinois, Urbana, Illinois 61801, United States

[2]Institute of Life Sciences, Swansea University Medical School, Swansea SA2 8PP, Wales, United Kingdom

[3]Departments of Biochemistry and Chemistry, University of Illinois, Urbana, Illinois 61801, United States

*Corresponding Author

Abstract

The continuous expansion of protein and genome sequence databases is an opportunity to identify novel enzymes with biotechnological applications. Whether applied to enzymology, chemical biology, systems biology, and microbiology, database mining must be "user-friendly" so that experimentalists can devise focused strategies to discover the *in vitro* activities and *in vivo* functions of uncharacterized enzymes. We developed a suite of genomic enzymology tools (https://efi.igb.illinois.edu/) to 1) generate sequence similarity networks (SSNs) for exploration of sequence-function space in protein families (EFI-EST) and 2) provide genome context for members of protein families (EFI-GNT). Integrated analysis of this complementary information allows to generate testable hypotheses about new functions. After a brief overview of EFI-EST and EFI-GNT, we describe selected applications that illustrate their use.

1 Introduction

When this Opinion was completed (August 1, 2020), the UniProt database contained 185,561,210 entries (Release 2020_03, June 22, 2020; https://www.uniprot.org/). The amount of information is "amazing": it should not be viewed as overwhelming but as an opportunity for discovery. The challenge is organizing and leveraging the data.

Most entries in the UniProt database (>80%) are assigned to at least one Pfam family and/or InterPro family that provides a (sometimes tentative) description of function. Sequence similarity networks (SSNs) are used widely for analyses of sequence-function space in protein families [1]. An SSN displays the results of an all-by-all pairwise sequence comparison (BLAST): each sequence is represented by a "node"; nodes are connected by a line ("edge") if they share a minimum user-specified sequence similarity. As the sequence similarity threshold increases, the nodes segregate into isofunctional clusters. Mapping experimentally established functions, e.g., SwissProt-curated, on the SSN allows identification of clusters with known functions;

uncharacterized clusters may contain enzymes in novel metabolic pathways and/or identifies the starting points for evolution of functions for novel applications.

In bacterial, archaeal, and fungal genomes, genes in operons and/or gene clusters are often functionally linked in a metabolic pathway. Thus, for an uncharacterized enzyme, genome context can provide information about the identity of the reaction as well as those of neighbors [2].

We use "genomic enzymology" [3] to describe the integration of analyses of sequence-function space in a protein family, together with the genome context of its members, to predict the enzymatic activities and the metabolic pathways in which they function. We developed a publicly accessible web resource (https://efi.igb.illinois.edu/) with tools that "democratize" genome enzymology [4-7]: 1) EFI-EST for generating SSNs (https://efi.igb.illinois.edu/efi-est/) and 2) EFI-GNT for mining genome contexts (https://efi.igb.illinois.edu/efi-gnt/). Since EFI-EST and EFI-GNT were introduced in 2014, >5,200 users have submitted >44,000 jobs to EFI-EST; >1,500 users have submitted >14,000 jobs to EFI-GNT. The tools have been cited in >300 publications; a list is available on the web resource "**? Training**" page (https://efi.igb.illinois.edu/training/).

We first provide brief descriptions of EFI-EST and EFI-GNT and then examples of their use for 1) surveying sequence-function space in protein families to identify candidate proteins with novel properties/functions and 2) discovering enzymes in novel metabolic pathways.

2 EFI-EST for Generating Sequence Similarity Networks (SSNs)

The Enzyme Function Initiative (EFI), a large-scale collaborative project supported by NIH/NIGMS (U54GM093342), developed strategies and tools to facilitate experimental assignment of *in vitro* activities and *in vivo* metabolic functions to uncharacterized enzymes discovered in genome projects [8]. The EFI-EST tool was developed for generating SSNs for protein families (https://efi.igb.illinois.edu/efi-est/). The user can specify a Pfam and/or InterPro family (using

Option B); the sequences are obtained from the UniProt database. For large families, EFI-EST uses either the UniRef90 or UniRef50 database in which sequences are conflated at 90% and 50% sequence identity threshold, respectively. The sequences contained in UniRef90 clusters almost certainly are orthologues; the sequences contained in UniRef50 clusters likely are orthologues. UniRef SSNs contain fewer nodes and edges than UniProt SSNs so can be more easily manipulated with laboratory laptop/desktop computers. The total set of accession IDs contained in individual clusters in UniRef SSNs can be used with Option C to obtain higher resolution SSNs. Alternatively, the user can use BLAST to collect sequences homologous to a query (Option A) to examine a localized high-resolution set of sequences within a protein family.

The user selects a minimum sequence similarity threshold (specified by an alignment score that is approximately the negative logarithm of the BLAST e-value) to draw edges to connect nodes, with the goal of segregating the dataset into isofunctional clusters of nodes. The SSN is visualized using Cytoscape, a platform for viewing complex networks; “node attributes” with taxonomic and bioinformatic information are provided. The node attributes, combined with genome context (next section), facilitate the choice of minimum alignment score threshold for generating isofunctional clusters in the SSN.

EFI-EST also provides the “Cluster Analysis” utility that provides easy access to multiple sequence alignments (MSA), WebLogos (based on the MSAs), hidden Markov models (HMMs; based on the MSAs), tables of residue conservation, and length histograms for each cluster in the input SSN. The MSAs and WebLogos provide an additional information for assessing sequence heterogeneity/isofunctionality within the clusters.

EFI-EST is a unique resource for generating and analyzing SSNs for protein families. As illustrated by the examples provided in Section 4, access to the SSNs for functionally diverse

protein (super)families facilitates identification of isofunctional clusters, including uncharacterized clusters that participate in undiscovered metabolic pathways.

3 EFI-GNT for Generating Genome Neighborhood Networks (GNNs) and Genome Neighborhood Diagrams (GNDs)

The EFI also developed the EFI-GNT tool to collect the genome neighborhoods of bacterial, archaeal, and fungal proteins in the clusters in an input SSN (https://efi.igb.illinois.edu/efi-gnt/). EFI-GNT provides two types of genome neighborhood networks (GNNs): 1) for each input SSN cluster, the identities and frequencies of co-occurrence of genome proximal Pfam families (with the descriptions providing information about possible functions), thereby assisting identification of the metabolic pathway in which the proteins in the SSN cluster function; and 2) for each genome proximal Pfam family, the identities of the query SSN clusters that identify the family, with this information allowing identification of orthologous SSN clusters. For each SSN cluster, a genome neighborhood diagram (GND) is provided for each bacterial, archaeal, and fungal protein, allowing visual inspection of the genome context. EFI-GNT also can be used to generate GNDs for homologues of a user-supplied sequence (using BLAST) or lists of user-supplied sequences.

EFI-GNT is a unique resource for retrieving, visualizing, and analyzing genome neighborhoods for user-supplied sequences. As illustrated by the examples provided in Section 5, this information is essential for predicting metabolic pathways and guiding the design of experimental approaches for functional verification.

4 Exploring Sequence-Function Space in SSNs to Identify New Functions and Pathways

The ability to visualize sequence-function in a family provides permits identification of both functionally characterized and uncharacterized (“unknown”) clusters. The “unknown” clusters

("dark matter") may contain enzymes with novel biological functions, unexpected mechanistic characteristics, and useful biotechnological applications. In this section, we describe several analyses of sequence-function space in families using SSNs.

4.1 Glycoside Hydrolases: GN128

Santos and coworkers used SSNs to explore sequence-function space in the GH128 glycoside hydrolase family; its members catalyze hydrolysis β-1,3-glycan linkages in polysaccharides [9]. An initial set of sequences was collected using characterized members as BLAST queries; a hidden Markov model (HMM) was constructed and used to identify additional homologues. The sequences were used to generate both a phylogenetic tree and an SSN (Figure 1A). Comparison of the tree and SSN illustrates the advantage of using an SSN to visualize sequence-function space: the isofunctional subgroups are easier to discern as emerging clusters in an SSN than the closely associated branches in a tree. In the tree, five of the seven subgroups are closely related phylogenetically and difficult to distinguish; in the SSN, the subgroups are well-defined as the result of the higher density of edges that connect members of isofunctional subgroups. Representative members of the subgroups were selected for structural, enzymological and molecular dynamics studies to characterize substrate specificity and mechanism, thereby informing the use of this family for biotechnological applications.

4.2 Diheme peroxidases

Elliott, Drennan and coworkers used an SSN to explore sequence-function space in Pfam family PF03150, diheme cytochrome c peroxidases (23K sequences in UniProt 2020_03) [10]. This family contains the abundant cytochrome c peroxidases (bCCPs) that catalyze reduction of H_2O_2 to H_2O in the cytoplasm of Gram-negative bacteria as well as MauG that catalyzes the mechanistically distinct oxidation/crosslinking of two Trp residues to generate the tryptophan

tryptophylquinone redox cofactor in methylamine dehydrogenase. The active sites of bCCP and MauG differ in the identities of the residues for the heme cofactor (Met-His in bCCP and Tyr-His in MauG), thereby explaining the different redox potentials.

In the SSN (Figure 1B), several additional clusters are observed, some segregated, some loosely associated with characterized proteins. This study focused on the emerging cluster IIIb that contains proteins encoded by *Bulkholderia* genomes that lack methylamine dehydrogenase; therefore, the sequences (designated BthA and BthB) were assumed to have novel functions. Characterization of BthA revealed that, like bCCP, it reduces H_2O_2 to H_2O but it also generates a bis-Fe(IV) species found on the MauG reaction coordinate. Although the physiological function is unknown, the encoding gene is adjacent to one for a purple-acid phosphatase; both genes are upregulated during anaerobic growth, suggesting functional linkage (post-translational modification of the phosphatase?). The SSN contains additional segregated and emerging clusters, suggesting that additional functions remain to be discovered in PF03150.

In a later study focusing on MbnH [11], an uncharacterized enzyme in PF03150 implicated in the biosynthesis of methanobactins [12], ribosomally-produced post-translationally modified copper-binding natural products, Keeney, Rosenzweig and coworkers described a further analysis of sequence-function space in PF03150 (Figure 1C). They structurally and spectroscopically characterized MbnH; although they were able to detect peroxidase activity, they were unable to specify a precise function, speculating MbnH also might be involved in post-translational modification of a genomically co-occurring partner protein, MbnP.

4.3 Rieske oxygenases

Rieske non-heme iron-dependent oxygenases catalyze C-H hydroxylation reactions in natural product biosynthesis, often cis-1,2-hydroxylation of aromatic substrates but also

monooxygenation reactions; these proteins contain a 2Fe-2S cluster ligated by two Cys and two His ligands. Bridwell-Rabb and Narayan generated the SSN for PF00355 (145K sequences in UniProt 2020_03), using an alignment score of 65 to segregate functions (Figure 1D) [13]. The SSN contains segregated clusters with functionally and structurally determined members. This study focused on structural characterization of SxtT and GstA located in the same SSN cluster, both catalyzing hydroxylation of dideoxysaxitoxin, albeit with different regiospecificities. Many of the larger clusters (light blue) in their SSN include aromatic ring hydroxylases. Inspection of their SSN reveals that large areas of sequence-function space are unexplored in PF00355.

4.4 Nitrogenase Superfamily

Ghenbreamlak and Mansoorabadi used an SSN to investigate functional diversity in the nitrogenase component 1 type oxidoreductase superfamily (PF00148; 25K sequences in UniProt 2020_03) [14]. The members of the superfamily couple the hydrolysis of ATP to the reduction of their substrates. The characterized members include nitrogenase (hydrolysis of 16 ATP to accomplish the six-electron reduction of N_2 to two molecules of ammonia and the two-electron reduction of two protons to H_2), dark-operative protochlorophyllide oxidoreductase (DPOR) and bacteriochlorophyll chlorophyllide *a* oxidoreductase (COR) that catalyze successive ATP-dependent two-electron reductions of intermediates in the biosynthesis of bacteriochlorophyll, and CfbCD that catalyzes the ATP-dependent six-electron reduction of an intermediate in the biosynthesis of coenzyme F430 in methanogenesis. The SSN (Figure 1E) segregates nitrogenases (Groups I, II, and III) from the coenzyme F430 (Group IV, CfbD) and bacteriochlorophyll (Group V, Bch) reductases. Notably, the SSN contains several clusters (A-A' through E-E') with uncharacterized functions. The genome neighborhoods of these clusters

suggest their involvement in the assembly of metalloclusters and/or the synthesis of a substrate for divergent homologues of nitrogenase.

4.5 2-Hydroxyacyl-CoA Dehydratase Superfamily

Similar to nitrogenase, members of the 2-hydroxyacyl-CoA dehydratase family (PF06050) use an ATP-dependent activator protein (from PF01869) to accomplish one-electron reduction reactions; the first characterized reactions involve formation of a ketyl radical of a thioester substrate [15]. If the substrate is a 2-hydroxyacyl-CoA, the ketyl radical undergoes dehydration to form the 2-enoyl-CoA; if the substrate is benzoyl-CoA, a second ATP-dependent one-electron reduction produces cyclohexa-1,5-dienoyl-CoA that can be catabolized. In yet unpublished work, we discovered members of PF06050 that catalyze the ATP-dependent four-electron reduction of $preQ_0$, a nitrile, to $preQ_1$, a primary amine, in queuosine biosynthesis, a reaction that does not use a thioester substrate.

Jeoung and Dobbek generated an SSN for PF06050 [16,17]. Using an alignment score (20; Figure 2A) that does not segregate the nodes into isofunctional clusters (95; unpublished), the SSN segregated the family based on sequence length and the number of Cys residues involved in coordination of the Fe-S cluster(s). Some clusters, e.g., the dehydratases and benzoyl-CoA reductases, contain sequences with three Cys residues that coordinate one 4Fe-4S cluster. Another SSN cluster contains sequences with seven conserved Cys residues that coordinate an 8Fe-9S double cubane cluster.

A member of the latter cluster from *Carboxydothermus hydrogenoformans* Z-290 ($DCCP_{ch}$) was characterized: unlike the heterodimeric hydratases and benzoyl-CoA reductase but like the $preQ_0$ reductase, it is encoded by a single gene; it also is encoded by an operon with an ATP-dependent activator/reductase (DCCP-R_{ch}). $DCCP_{ch}$ catalyzes the ATP-dependent reduction of acetylene (using dithiothreitol as the source of electrons). The structure was determined, providing the

structure of an unusual double cubane 8Fe-9S cluster. The identity of the substrate and physiological reaction is unknown.

4.6 Flavin-dependent amine oxidases

Several families of flavin-dependent oxidases have been described, including the large FAD-dependent oxidoreductase family (PF01266; 268K sequences in UniProt 2020_03) with both amine and alcohol dehydrogenases and the smaller flavin-dependent amine oxidase family (PF01593; 123K sequences in UniProt 2020_03) that includes the extensively characterized monoamine oxidase (MAO). Tararina and Allen generated the SSN for PF01593 to guide investigations of the structural bases for specificity, with the goal of (re)engineering specificity (Figure 2B) [18]. Members of the family share a bidomain structure, a flavin-binding Rossman fold domain and a substrate-binding domain with hotdog-like fold and helical bundle subdomains. They mapped the known functions to the SSN and segregated it into clusters/subgroups with different specificities. The subgroups differ in the sequences/structures of the substrate-binding domain; these were associated with changes in the entrance cavities to the active sites. They also used EFI-GNT to determine the genome contexts for the subgroups. As is typical for large families, the SSN contains many uncharacterized clusters, guiding future studies that would contribute to a more detailed understanding of the sequence-structure changes involved in the evolution of substrate specificity.

4.7 Flavin-Dependent Aromatic Halogenases

Lewis and coworkers used SSNs to survey substrate specificity space in the flavin-dependent halogenase family, with the goal of using “family-wide activity profiling” to identify candidates for

halogenation of unnatural substrates [19]. They proposed that sampling sequence-function space in the SSN is a better approach than directed evolution for accessing halogenases for synthetic applications. Instead of using Option B (families) to generate the SSN (PF04820, tryptophan halogenase; 15K sequences in UniProt 2020_03), the sequence of RebH, a 7-chlorotryptophanase in the biosynthetic pathway for rebeccamycin, was used as the query for Option A (BLAST). Three large clusters were identified using a minimum sequence similarity threshold of 30% (alignment score 70; Figure 2C, top panel/Level 1) to draw edges); mapping of the characterized functions identified clusters specific for indole, phenol, or pyrrole substrates. Uncharacterized clusters were also present. Using a minimum sequence similarity threshold of 40% (alignment score 140; Figure 2C, bottom panel/Level 2), the indole and phenol clusters further segregated into clusters with distinct substrate specificities. A set of diverse proteins were expressed, purified, and subjected to substrate profiling with a library of substrates. As highlighted in an accompanying Commentary [20], this study provides an instructive example of the use of SSNs for identification of candidate enzymes for synthetic potential (or, more generally, proteins with desired chemical and/or physical properties).

4.8 Flavin-Dependent Amino Acid Halogenases

Chang and coworkers used an SSN to investigate the substrate specificities of amino acid-halogenases that are members of a novel Fe(II)/α-ketoglutarate-dependent family within the extremely large cupin superfamily [21]. They had discovered BesD, a novel 4-chloro-Lys halogenase in the pathway for alkyne-containing amino acids in *Streptomyces cattleya*. The sequence of BesD was used as the BLAST query to identify homologues. When the SSN was generated using a minimum alignment score threshold of 88 (Figure 2D), the sequences segregated into isofunctional clusters that were subjected to functional screening. The sequences were also analyzed using EFI-GNT, with the genome neighborhoods (GNDs) containing genes encoding enzymes involved in amino acid metabolism.

5 Using Genome Context in GNNs and GNDs to Identify New Functions and Pathways

Analysis of sequence-function space in an SSN identifies clusters with uncharacterized functions; if the family is curated by Pfam or the other family databases integrated in InterPro, the type of chemical reaction catalyzed by an uncharacterized cluster may be predicted. However, the precise *in vitro* activity and *in vivo* function likely requires knowledge of the types of reactions catalyzed by functionally linked proteins in metabolic pathways. For bacterial, archaeal, and fungal proteins, these often can be identified from analyses of genome neighborhood context, because metabolic pathways frequently are encoded by proximal genes (operons and/or gene clusters). In this section we provide examples of the use of EFI-GNT to provide genome context for uncharacterized enzymes, thereby allowing the *in vitro* activities and *in vivo* functions to be experimentally characterized

5.1 Catabolic Pathways for D-Apiose, a Branched Pentose in Plant Cell Walls

Even with knowledge of the types of reactions in an uncharacterized catabolic pathway, the identities of the substrates and products and, therefore, the exact sequence of reactions in the pathway usually cannot be specified without knowing the identity of the substrate for the first enzyme. In bacteria, transport systems that import an extracellular solute for catabolism often are encoded by a genome neighborhood that also encodes the catabolic pathway. In a large-scale program that used ThermoFluor screening and a large physical library to identify ligands for solute binding proteins (SBPs) for ABC transport systems [22], three SBPs from Pfam PF13407 were identified that bind D-apiose [23], a branched pentose found in the rhamnogalacturonan-II component of plant cell walls; these SBPs were located in two SSN clusters.

The genome neighborhoods of the D-apiose-binding SBPs were identified using SSNs and GNNs. Two types of catabolic pathways were identified (Figure 3): 1) a non-oxidative pathway using a transketolase to convert the branched D-apiose and D-glyceraldehyde 3-phosphate to D-xylulose 5-phosphate and dihydroxyacetone phosphate, intermediates in the pentose phosphate and glycolysis pathways, respectively (Figure 3A); and 2) three oxidative pathways initiated by oxidation/ring-opening of D-apiofuranose to D-apionate followed by an unprecedented oxidation/isomerization reaction involving migration of a hydroxymethyl group to generate “3-oxo-isoapioniate” (Figure 3B). The non-oxidative pathway is present in species of Bacteroides found in the human gut microbiome; the nonoxidative pathways are found in bacteria in other ecological niches, e.g., soil.

5.2 A Novel Catabolic Pathway for L-Ascorbate

Stack and coworkers used SSNs and GNNs to guide discovery of a novel bacterial pathway for L-ascorbate catabolism in *Ralstonia eutropha* H16 (*Cupriavidus necator* ATCC17699) [24] (Figure 4, panels A-C); the same pathway, or segments of the pathway, were identified in hundreds of additional bacteria species. Two similar catabolic pathways, one aerobic and one anaerobic [25], previously had been characterized in *Escherichia coli* K-12 and then found in many other bacteria. *R. eutropha* can utilize L-ascorbate as sole carbon source, but does not encode the *E. coli* pathways. RNA-Seq was used to identify genes up-regulated during growth on L-ascorbate; 11 genes, located in three distinct gene clusters/modules, were identified. SSNs were generated for the protein families represented in these modules; GNNs and GNDs then were generated to guide the discovery of the pathway encoded by the three modules. Module 1 (Figure 4A) encodes an L-ascorbate oxidase to generate dehydro-L-ascorbate (a lactone) and a lactonase to generate 2,3-diketo-L-gulonate (DKG). Module 2 (Figure 4B) encodes a DKG mutase that catalyzes an unprecedented benzylic acid rearrangement to generate 2-carboxy-L-lyxonolactone and a lactonase to generate 2-carboxy-L-lyxonate (Clx). Module 3 (Figure 4C) encodes a novel NAD^+-

dependent decarboxylase that converts Clx to L-lyxonate as well as orthologues of three enzymes in a previously characterized catabolic pathway for L-lyxonate that produces α-ketoglutarate [26].

5.3 Queuine Salvage Pathways

Yuan and coworkers used SSNs and GNNs to identify bacterial salvage pathways for queuine, a precursor of the modified base queuosine found in tRNAs [27]. Most eubacteria can synthesize queuosine *de novo*, via $preQ_0$ and $preQ_1$ intermediates, in a metabolic pathway of eight enzymes starting with GTP; eukaryotes and some bacteria, including pathogens such as *Chlamydia* cannot synthesize queuosine so must salvage the queuine precursor. For salvage, most members of the YhhQ transporter family import $preQ_0$ and $preQ_1$, with a tRNA guanine transglycosylase (TGT) exchanging a guanine with $preQ_1$ that is further modified to generate queuosine.

The SSN was generated for the YhhQ family. EFI-GNT was used to identify clusters with proximal TGT homologues but without biosynthetic genes for queuosine (Figure 5A). Focusing on species of *Chlamydia* that can salvage queuine but not $preQ_1$, a multiple sequence alignment of its TGT together with homology models suggested an enlarged active site that can accommodate the bulkier queuine moiety, thereby explaining the structural basis for queuine salvage.

This study also elucidated the salvage pathway by which *Clostridioides difficile* can use queuosine as a source of salvageable queuine (Figure 5B). A transporter specific for queuosine was identified that is under the control of a $preQ_1$ riboswitch. The gene encoding the transporter is located in a three-gene operon that also encodes a nucleoside hydrolase and a member of the radical SAM superfamily. Enzymological characterization of the nucleoside hydrolase demonstrated that it is specific for queuosine, generating queuine; characterization of the radical SAM enzyme demonstrated that it is a lyase that frees the dihydroxycyclopentene moiety from queuine, liberating $preQ_1$ that can be incorporated into tRNA with TGT and then modified to generate the queuosine nucleotide.

5.4 Pathways for Organosulfur Catabolism

Balskus and coworkers generated the SSN for the glycyl radical enzyme (GRE) superfamily (IPR004184) and used this to 1) describe sequence-function space in the superfamily and 2) develop chemically guided functional profiling (CGFP) that maps metagenome abundance to isofunctional clusters (families) in the SSN [28]. In subsequent independent studies, Peck, Balskus, and coworkers [29] and Zhao and coworkers [30] identified a GRE involved in the catabolic pathway for taurine (2-aminosulfonate) in *Bilophila wadsworthia* found in the human gut microbiome. Isethionate sulfite lyase cleaves taurine to acetaldehyde and sufite, the former used to generate acetyl-CoA and the latter used as an electron acceptor with formation of sulfide (Figure 6A).

Zhao and coworkers also discovered two members of the GRE superfamily that degrade dihydropropanesulfonate (S-DHPS), a microbial degradation product of 6-sulfo-D-glucopyranose (sulfoquinovose) that is a component of an abundant plant sulfolipid [31]. Considering the structure of S-DHPS, a member of the GRE superfamily could catalyze either 1) desulfuration to produce hydroxyacetone and sufite by HpsG, analogous to the reaction catalyzed by taurine lyase or 2) dehydration to produce 3-sulfopropionaldehyde by HpfG, analogous to the reactions catalyzed by glycerol dehydration and 1,2-propanediol dehydrate that members of the GRE superfamily (Figure 6B). SSNs and GNNs were used to confirm the identities of these pathways: the desulfuration pathway involves a GRE homologous to isethionate sulfate lyase (HpsG); the dehydration pathway involves a GRE homologous to both glycerol and 1,3-propanediol dehydratases (HpfG).

5.5 A Pathway for Herbicide Degradation

Cicchillo and workers used SSNs and GNNs to identify aryloxyalkanoate dioxygenases (AADs) for engineering herbicide resistant transgenic plants [32]. TfdA that degrades the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D) (Figure 7A) is homologous with bacterial taurine dioxygenases, suggesting AADs have promiscuous substrate specificities. The hypothesis was that divergent TfdAs that would be candidates for degrading herbicides structurally related to 2,4-D and that these would be located in operons/gene clusters in the known Tfd pathway.

The SSN of TfdA and bacterial homologues collected using BLAST was generated with EFI-EST; the SSN was used as the input for EFI-GNT to retrieve genome neighborhood context. EFI-GNT adds a node attribute to Color SSNs generated with the Color SSN utility that includes the Pfam/InterPro family identifiers of genome neighbors, allowing easy identification of nodes that are genome proximal to enzymes in the Tfd pathway (Figure 7B). The SSN with nodes colored according to the identities of genome neighbors is shown in **Figure 3**. Fifty-nine new dioxygenases were identified that were screened for herbicide degradation. The synergistic use of EFI-EST and EFI-GNT “simplified” the discovery of these AADs.

6 Identification and analysis of biosynthetic gene clusters

Although not the focus of this Opinion, several web tools are widely used for the discovery of biosynthetic gene clusters (BGCs), including antiSMASH [33,34], PRISM [35,36], and RODEO [37]. For antiSMASH and PRISM, user-specified sequenced genomes are used as the input; the output identifies the encoded BGCs. For RODEO, user-supplied lists of enzymes involved in the syntheses of ribosomally synthesized and post-translationally modified peptides (RiPPs) are used as input; the retrieved genome neighborhoods are mined for short precursor peptides that often are not identified in genome annotation.

EFI-EST and EFI-GNT have been used to identify genome neighborhoods that contain orthologues/homologues of key enzymes in natural product biosynthesis, e.g., thioamide synthases [38], enzymes involved in diazeniumdiolate siderophore synthesis [39], enzymes in endopyrrole biosynthesis [40], enzymes in polytheonamide biosynthesis [41], and flavoprotein monooxygenases (Baeyer-Villiger-type oxygen insertion) in polyketide synthase natural products [42]. In some cases, Option A of EFI-EST (BLAST using a user-specified query) is used to generate the SSN of homologues/orthologues in the UniProt database; EFI-GNT is used to visualize/analyze the BGCs. In other cases, homologues/orthologues identified by other approaches are used with EFI-GNT to visualize/analyze the BGCs.

7 Summary and Outlook

SSNs generated with EFI-EST and GNNs/GNDs generated with EFI-GNT are used by the enzymology, chemical biology, and microbiology communities to 1) identify “dark matter” targets in protein families for functional characterization (SSNs) and 2) provide functional context information to guide experimental verification of their *in vitro* activities and *in vivo* functions (GNNs and GNDs). In addition to leveraging the discovery of novel metabolic pathways, an SSN provides the ability to survey and broadly sample sequence-function space in a protein family for novel enzymatic activities and/or physical properties. The publications that describe their use, only a small number of which are highlighted in this Opinion, can be used to stimulate additional applications. A complete list of publications is available on the web resource “**? Training**” page (https://efi.igb.illinois.edu/training/).

The genomic enzymology tools described in this Opinion were developed with support from NIGMS, initially a large-scale collaborative project (U54GM093342; EFI) focused on the development of strategies and tools for assigning functions to uncharacterized enzymes discovered in genome projects and currently a Program Project (P01GM118303) focused on the

discovery of novel metabolic pathways using the ligand specificities of transport system solute binding proteins to identify both the substrate for the first enzyme in the pathway and the genes encoding the enzymes in the metabolic pathway (genome neighbors of the transport system). However, dedicated support for support of bioinformatics/genomic enzymology resources such as EFI-EST and EFI-GNT is difficult to secure, i.e., the resource does not contribute original research but provides tools that facilitate research projects. As we noted in a recent Current Opinion in Chemical Biology [6]: “We hope that members of the community and, therefore, funding agencies will recognize the need to support publicly accessible genomic enzymology resources for leveraging sequence databases. The required investment is small compared to both the cost and potential impact of genome projects.”

Conflict of interest statement

Nothing declared.

Acknowledgements

The web resource that provides “democratized” community access to EFI-EST and EFI-GNT was originally developed with support from NIH U54GM093342 and currently is supported by NIH P01GM118303. R.Z. is currently supported by the European Union’s Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowships Grant agreement H2020-MSCA-IF-2018 839116 deCrYPtion.

7 References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

** of outstanding interest

1. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC: **Using sequence similarity networks for visualization of relationships across diverse protein superfamilies**. *PLoS One* 2009, **4**:e4345.

* The description of SSNs that demonstrated their importance in analyzing sequence-function space in protein famililes to the enzymology community.

2. Zhao S, Sakai A, Zhang X, Vetting MW, Kumar R, Hillerich B, San Francisco B, Solbiati J, Steves A, Brown S, et al.: **Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks**. *Elife* 2014, **3**.

3. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies**. *Annu Rev Biochem* 2001, **70**:209-246.

4. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, Whalen KL: **Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks**. *Biochim Biophys Acta* 2015, **1854**:1019-1037.

5. Gerlt JA: **Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence-Function Space and Genome Context to Discover Novel Functions**. *Biochemistry* 2017, **56**:4293-4308.

6. Zallot R, Oberg NO, Gerlt JA: **'Democratized' genomic enzymology web tools for functional assignment**. *Curr Opin Chem Biol* 2018, **47**:77-85.

7. Zallot R, Oberg N, Gerlt JA: **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019, **58**:4169-4182.

** A detailed description of the "democratized" EFI genomic enzymology web resource that includes EFI-EST, EFI-GNT, and EFI-CGFP.

8. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W, et al.: **The Enzyme Function Initiative**. *Biochemistry* 2011, **50**:9950-9962.

9. Santos CR, Costa P, Vieira PS, Gonzalez SET, Correa TLR, Lima EA, Mandelli F, Pirolla RAS, Domingues MN, Cabral L, et al.: **Structural insights into beta-1,3-glucan cleavage by a glycoside hydrolase family**. *Nat Chem Biol* 2020.

** The phylogenetic tree and SSN for the GH128 family are compared; the subgroups are easier to discern in the SSN than in the tree.

10. Rizzolo K, Cohen SE, Weitz AC, Muñoz MML, Hendrich MP, Drennan CL, Elliott SJ: **A widely distributed diheme enzyme from Burkholderia that displays an atypically stable bis-Fe (IV) state**. *Nature communications* 2019, **10**:1101.

** An exploration of sequence-function space in the diheme cytochrome c peroxidase superfamily; an uncharacterized cluster was selected for spectroscopic characterization.

11. Kenney GE, Dassama LMK, Manesis AC, Ross MO, Chen S, Hoffman BM, Rosenzweig AC: **MbnH is a diheme MauG-like protein associated with microbial copper homeostasis**. *J Biol Chem* 2019, **294**:16141-16151.

12. Kenney GE, Dassama LMK, Pandelia ME, Gizzi AS, Martinie RJ, Gao P, DeHart CJ, Schachner LF, Skinner OS, Ro SY, et al.: **The biosynthesis of methanobactin**. *Science* 2018, **359**:1411-1416.

13. Lukowski AL, Liu J, Bridwell-Rabb J, Narayan ARH: **Structural basis for divergent C-H hydroxylation selectivity in two Rieske oxygenases**. *Nat Commun* 2020, **11**:2991.

** An exploration of sequence-function space in the Rieske non-heme iron-dependent oxygenase superfamily

14. Ghebreamlak SM, Mansoorabadi SO: **Divergent Members of the Nitrogenase Superfamily: Tetrapyrrole Biosynthesis and Beyond**. *Chembiochem* 2020.

An exploration of sequence-function space in the nitrogenase superfamily.

15. Buckel W: **Enzymatic Reactions Involving Ketyls: From a Chemical Curiosity to a General Biochemical Mechanism**. *Biochemistry* 2019, **58**:5221-5233.

16. Jeoung J-H, Dobbek H: **ATP-dependent substrate reduction at an [Fe8S9] double-cubane cluster**. *Proceedings of the National Academy of Sciences* 2018, **115**:2994-2999.

** An exploration of sequence-function space in the 2-hydroxyacyl-CoA dehydratase superfamily; an uncharacterized cluster with a novel 8Fe-9Cys double cubane cluster was discovered

17. Jeoung JH, Martins BM, Dobbek H: **Double-Cubane [8Fe9S] Clusters: A Novel Nitrogenase-Related Cofactor in Biology**. *Chembiochem* 2020.

18. Tararina MA, Allen KN: **Bioinformatic Analysis of the Flavin-Dependent Amine Oxidase Superfamily: Adaptations for Substrate Specificity and Catalytic Diversity**. *J Mol Biol* 2020.

* An exploration of sequence-function space in the flavin-dependent amine oxidase family.

19. Fisher BF, Snodgrass HM, Jones KA, Andorfer MC, Lewis JC: **Site-Selective C-H Halogenation Using Flavin-Dependent Halogenases Identified via Family-Wide Activity Profiling**. *ACS Cent Sci* 2019, **5**:1844-1856.

* An exploration of sequence-function space in the flavin-dependent halogenase family; the SSN segregated the family into clusters specific for different types of aromatic substrates.

20. Rodriguez Benitez A, Narayan ARH: **Frontiers in Biocatalysis: Profiling Function across Sequence Space**. *ACS Cent Sci* 2019, **5**:1747-1749.

21. Neugebauer ME, Sumida KH, Pelton JG, McMurry JL, Marchand JA, Chang MCY: **A family of radical halogenases for the engineering of amino-acid-based products**. *Nature chemical biology* 2019, **15**:1009-1016.

22. Vetting MW, Al-Obaidi N, Zhao S, San Francisco B, Kim J, Wichelecki DJ, Bouvier JT, Solbiati JO, Vu H, Zhang X, et al.: **Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes**. *Biochemistry* 2015, **54**:909-931.

23. Carter MS, Zhang X, Huang H, Bouvier JT, Francisco BS, Vetting MW, Al-Obaidi N, Bonanno JB, Ghosh A, Zallot RG, et al.: **Functional assignment of multiple catabolic pathways for D-apiose**. *Nat Chem Biol* 2018, **14**:696-705.

** SSNs and GNNs are used to discover four novel pathways for D-apiose catabolism, one redox-neutral using a transketolase and three using a oxidoisomerase. This study is noteworthy because it demonstrates the iterative use of SSNs and GNNs to discover multiple catabolic pathways.

24. Stack TMM, Morrison KN, Dettmer TM, Wille B, Kim C, Joyce R, Jermain M, Naing YT, Bhatti K, Francisco BS, et al.: **Characterization of an l-Ascorbate Catabolic Pathway with Unprecedented Enzymatic Transformations**. *J Am Chem Soc* 2020, **142**:1657-1661. * SSNs and GNNs are used to discover a novel pathway for L-ascorbate catabolism.

25. Yew WS, Gerlt JA: **Utilization of L-ascorbate by Escherichia coli K-12: assignments of functions to products of the yjf-sga and yia-sgb operons**. *J Bacteriol* 2002, **184**:302-306.

26. Ghasempur S, Eswaramoorthy S, Hillerich BS, Seidel RD, Swaminathan S, Almo SC, Gerlt JA: **Discovery of a novel L-lyxonate degradation pathway in Pseudomonas aeruginosa PAO1**. *Biochemistry* 2014, **53**:3357-3366.

27. Yuan Y, Zallot R, Grove TL, Payan DJ, Martin-Verstraete I, Sepic S, Balamkundu S, Neelakandan R, Gadi VK, Liu CF, et al.: **Discovery of novel bacterial queuine salvage**

enzymes and pathways in human pathogens. *Proc Natl Acad Sci U S A* 2019, **116**:19126-19135.

* SSNs and GNNs are used to discover two pathways for queuine salvage, one containing a novel radical SAM enzyme.

28. Levin BJ, Huang YY, Peck SC, Wei Y, Martinez-Del Campo A, Marks JA, Franzosa EA, Huttenhower C, Balskus EP: **A prominent glycyl radical enzyme in human gut microbiomes metabolizes trans-4-hydroxy-l-proline**. *Science* 2017, **355**.

29. Peck SC, Denger K, Burrichter A, Irwin SM, Balskus EP, Schleheck D: **A glycyl radical enzyme enables hydrogen sulfide production by the human intestinal bacterium Bilophila wadsworthia**. *Proc Natl Acad Sci U S A* 2019, **116**:3171-3176.

** SSNs and GNNs are used to discover a pathway for taurine/isethionate catabolism in the human gut microbiome.

30. Tong Y, Wei Y, Hu Y, Ang EL, Zhao H, Zhang Y: **A Pathway for Isethionate Dissimilation in Bacillus krulwichiae**. *Appl Environ Microbiol* 2019, **85**:AEM. 00793-00719.

** SSNs and GNNs are used to discover a pathway for taurine/isethionate catabolism in the human gut microbiome.

31. Liu J, Wei Y, Lin L, Teng L, Yin J, Lu Q, Chen J, Zheng Y, Li Y, Xu R, et al.: **Two radical-dependent mechanisms for anaerobic degradation of the globally abundant organosulfur compound dihydroxypropanesulfonate**. *Proc Natl Acad Sci U S A* 2020, **117**:15599-15608.

32. Chekan JR, Ongpipattanakul C, Wright TR, Zhang B, Bollinger JM, Jr., Rajakovich LJ, Krebs C, Cicchillo RM, Nair SK: **Molecular basis for enantioselective herbicide degradation imparted by aryloxyalkanoate dioxygenases in transgenic plants**. *Proc Natl Acad Sci U S A* 2019, **116**:13299-13304.

** SSNs and GNNs are used to discover dioxygenases that can be used to engineer herbicide resistant plants.

33. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R: **antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences**. *Nucleic Acids Res* 2011, **39**:W339-346.

34. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T: **antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline**. *Nucleic Acids Res* 2019, **47**:W81-W87.

35. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, Wyatt MA, Magarvey NA: **Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM)**. *Nucleic Acids Res* 2015, **43**:9645-9662.

36. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA: **PRISM 3: expanded prediction of natural product chemical structures from microbial genomes**. *Nucleic Acids Res* 2017, **45**:W49-W54.

37. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai HC, Zakai UI, Mitchell DA: **A new genome-mining tool redefines the lasso peptide biosynthetic landscape**. *Nat Chem Biol* 2017, **13**:470-478.

38. Dunbar KL, Dell M, Molloy EM, Kloss F, Hertweck C: **Reconstitution of Iterative Thioamidation in Closthioamide Biosynthesis Reveals Tailoring Strategy for Nonribosomal Peptide Backbones**. *Angew Chem Int Ed Engl* 2019, **58**:13014-13018.

** This and the following references illustrate the use of EFI-GNT to generate GNDs to leverage discovery of biosynthetic gene clusters.

39. Hermenau R, Mehl JL, Ishida K, Dose B, Pidot SJ, Stinear TP, Hertweck C: **Genomics-Driven Discovery of NO-Donating Diazeniumdiolate Siderophores in Diverse Plant-Associated Bacteria**. *Angew Chem Int Ed Engl* 2019, **58**:13024-13029.

40. Niehs SP, Dose B, Scherlach K, Pidot SJ, Stinear TP, Hertweck C: **Genome Mining Reveals Endopyrroles from a Nonribosomal Peptide Assembly Line Triggered in Fungal-Bacterial Symbiosis**. *ACS Chem Biol* 2019, **14**:1811-1818.

41. Bosch NM, Borsa M, Greczmiel U, Morinaka BI, Gugger M, Oxenius A, Vagstad AL, Piel J: **Landornamides: Antiviral Ornithine-Containing Ribosomal Peptides Discovered through Genome Mining**. *Angew Chem Int Ed Engl* 2020, **59**:11763-11768.

42. Ueoka R, Meoded RA, Gran-Scheuch A, Bhushan A, Fraaije MW, Piel J: **Genome Mining of Oxidation Modules in trans-Acyltransferase Polyketide Synthases Reveals a Culturable Source for Lobatamides**. *Angew Chem Int Ed Engl* 2020, **59**:7761-7765.

Legends to Figures

Figure 1. Sequence similarity networks for protein families discussed in the text. Panel A (section 4.1), glycoside hydrolase family GH128 [9]. Reprinted with permission from *Nature Chemical Biology*. Panel B (section 4.2), diheme cytochrome c peroxidase family [10]. Reprinted with permission from *Nature Communications*. Panel C (section 4.2), diheme cytochrome c peroxidase family [11]. Reprinted with permission from *Journal of Biological Chemistry*. Panel D (section 4.2), Rieske non-heme iron-dependent oxygenases [13]. Reprinted with permission from *Nature Communications*. Panel E (section 4.4), nitrogenase component 1 type oxidoreductase superfamily [14]. Reprinted with permission from *ChemBioChem*.

Figure 2. Sequence similarity networks for protein families discussed in the text. Panel A (section 4.5), 2-hydroxyacyl-CoA dehydratase superfamily [16,17]. Reprinted with permission from *Proceedings of the National Academy of Sciences USA*. Panel B (section 4.6), flavin-dependent amine oxidases [18]. Reprinted with permission from *Journal of Molecular Biology.* Panel C (section 4.7), flavin-dependent aromatic halogenases [19]. Reprinted with permission from *Journal of the American Chemical Society*. Panel D (section 4.8), flavin-dependent amino acid halogenases [21]. Reprinted with permission from *Nature Chemical Biology*.

Figure 3. Catabolic pathways for D-apiose. (Section 5.1). Panel A, nonoxidative pathway involving a transketolase. Panel B, oxidative pathways.

Figure 4. Catabolic pathway for L-ascorbate. (Section 5.2). Reprinted with permission from *Journal of the American Chemical Society*.

Figure 5. Salvage pathways for queuine (Section 5.3). Panel A, sequence similarity network for the YhhQ family of transporters, colored to show genome neighborhood context. Panel B, salvage pathway for queuine involving a queuosine hydrolase and a queuine lyase to generate $preQ_1$. Reprinted with permission from *Proceedings of the National Academy of Sciences USA*.

Figure 6. C-S bond cleavage reactions catalyzed by members of the glycyl radical enzyme (GRE) superfamily (Section 5.4). Panel A, catabolism of taurine to acetyl-CoA. Panel B, two catabolic pathways for dihydropropanesulfonate (S-DHPS), a microbial degradation product of 6-sulfo-D-glucopyranose (sulfoquinovose). Reprinted with permission from *Proceedings of the National Academy of Sciences USA*.

Figure 7. Identification of arylalkanoate dioxygenases (AADs) for engineering herbicide resistance (Section 5.5). Panel A, the reaction catalyzed by TfdA in the catabolic pathway for 2,4-D. Panel B, the SSN for selected members of the TauD superfamily with nodes colored to identify TfdAs based on genome neighborhood context. Reprinted with permission from *Proceedings of the National Academy of Sciences USA*.

Figure 1

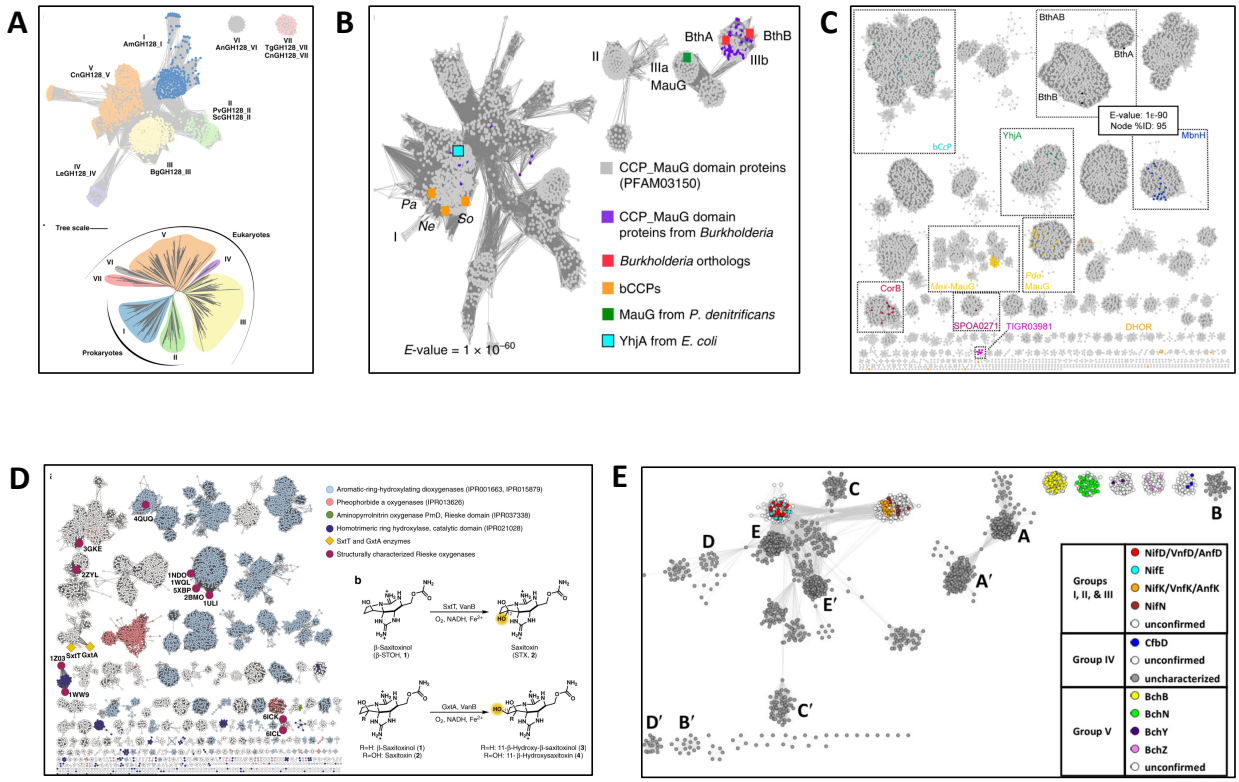

A
B
C
D
E
CCP_MauG domain proteins (PFAM03150)
CCP_MauG domain proteins from Burkholderia
Burkholderia orthologs
bCCPs
MauG from P. denitrificans
YhjA from E. coli
E-value = 1 × 10−60
BthA
BthB
MauG
Pa
Ne
So
E-value: 1e-90
Node %ID: 95
Groups I, II, & III
NifD/VnfD/AnfD
NifE
NifK/VnfK/AnfK
NifN
unconfirmed
Group IV
CfbD
unconfirmed
uncharacterized
Group V
BchB
BchN
BchY
BchZ
unconfirmed

Figure 2

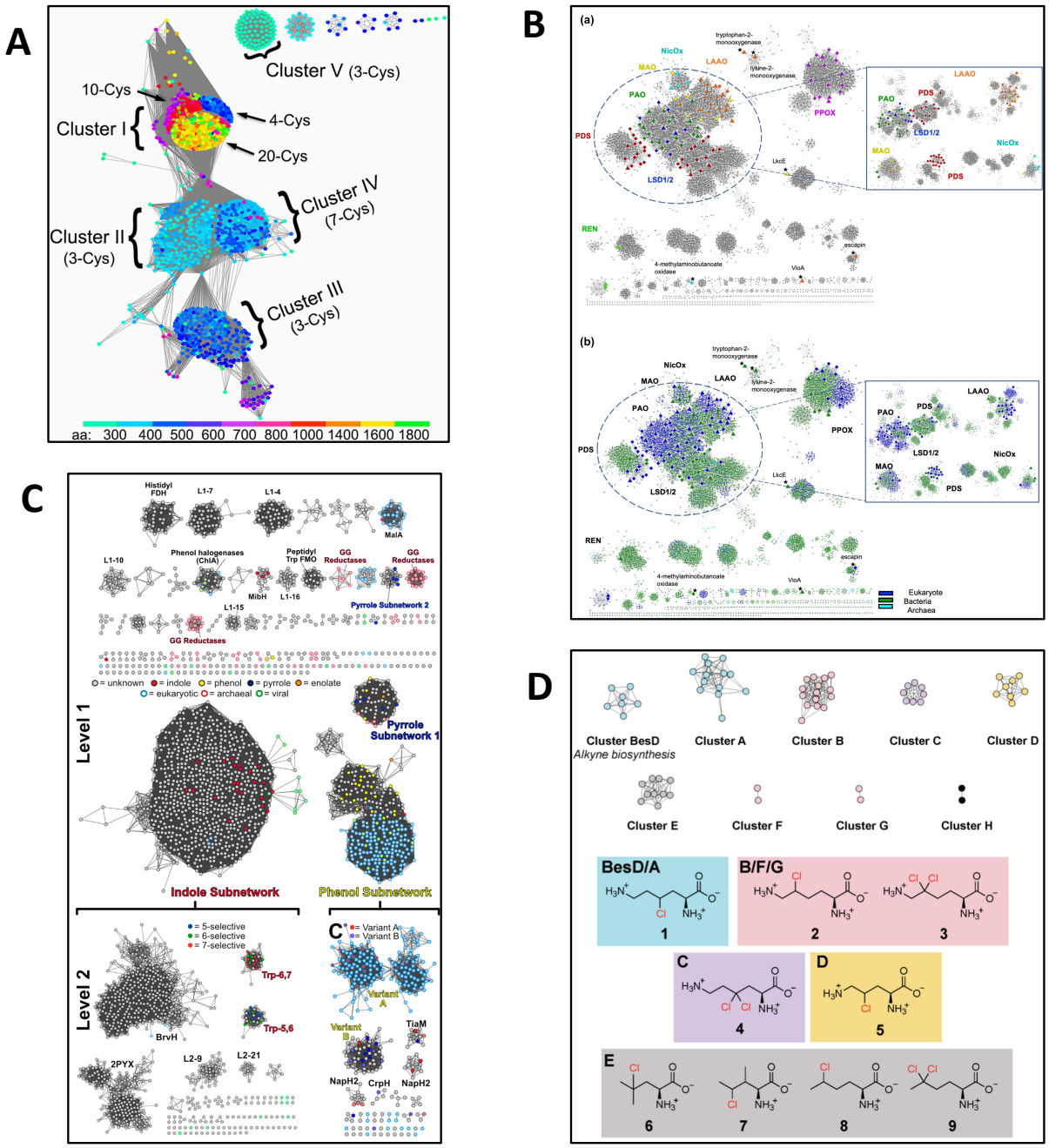

A
Cluster V (3-Cys)
10-Cys
Cluster I
4-Cys
20-Cys
Cluster IV
(7-Cys)
Cluster II
(3-Cys)
Cluster III
(3-Cys)
aa: 300 400 500 600 700 800 1000 1400 1600 1800
B
(a)
(b)
NicOx
MAO
LAAO
PAO
PDS
LSD1/2
PPOX
LkcE
REN
escapin
VioA
Eukaryote
Bacteria
Archaea
C
Histidyl FDH
L1-7
L1-4
MalA
L1-10
Phenol halogenases (ChlA)
MlbH
L1-16
Peptidyl Trp FMO
GG Reductases
L1-15
Pyrrole Subnetwork 2
= unknown = indole = phenol = pyrrole = enolate
= eukaryotic = archaeal = viral
Level 1
Pyrrole Subnetwork 1
Indole Subnetwork
Phenol Subnetwork
Level 2
= 5-selective
= 6-selective
= 7-selective
Trp-6,7
Trp-5,6
BrvH
2PYX
L2-9
L2-21
= Variant A
= Variant B
Variant A
Variant B
TiaM
NapH2
CrpH
D
Cluster BesD
Alkyne biosynthesis
Cluster A
Cluster B
Cluster C
Cluster D
Cluster E
Cluster F
Cluster G
Cluster H
BesD/A
B/F/G
1
2
3
4
5
6
7
8
9

Figure 3

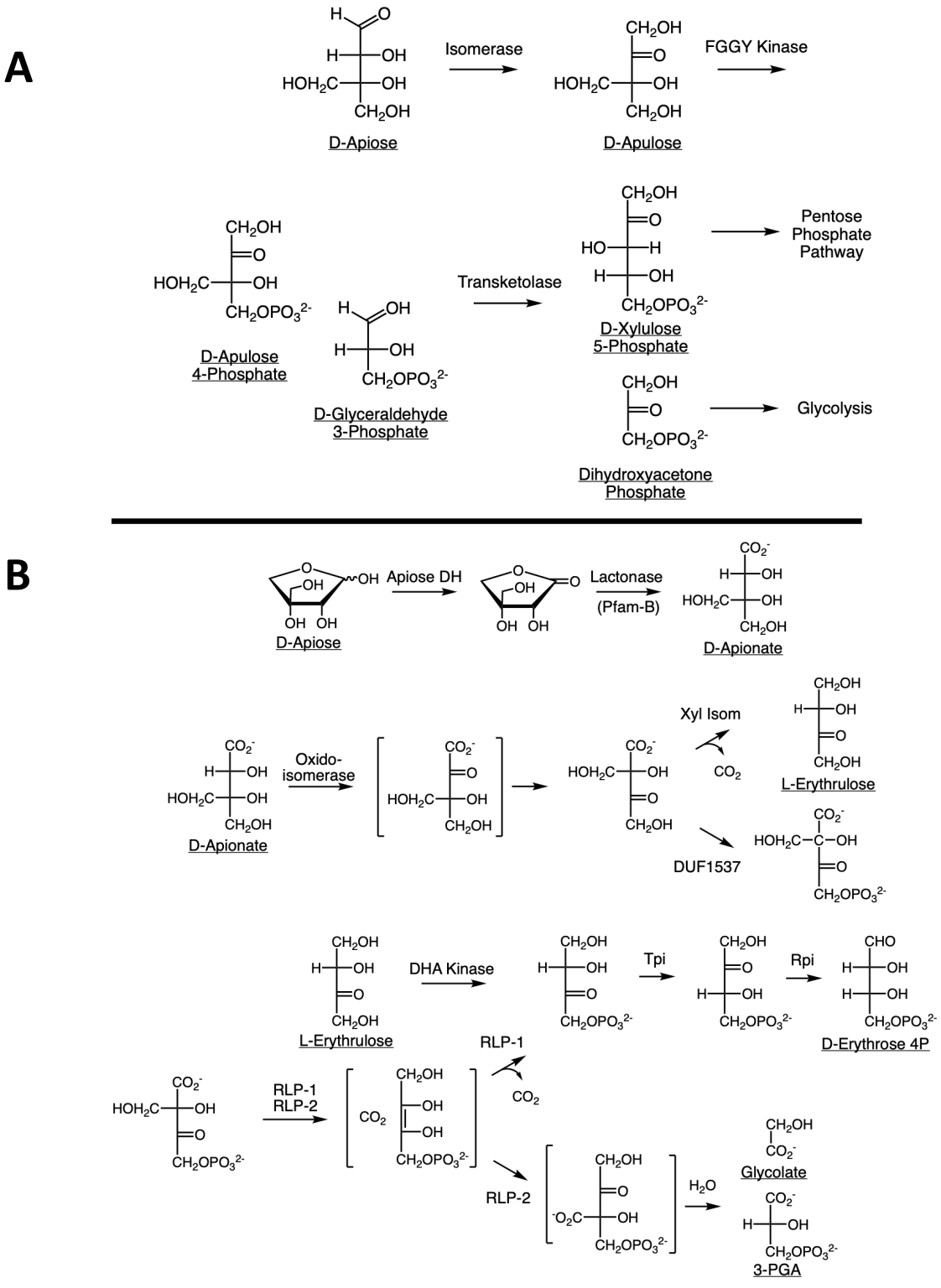

A
D-Apiose
Isomerase
D-Apulose
FGGY Kinase
D-Apulose 4-Phosphate
Transketolase
D-Glyceraldehyde 3-Phosphate
D-Xylulose 5-Phosphate
Pentose Phosphate Pathway
Dihydroxyacetone Phosphate
Glycolysis
B
D-Apiose
Apiose DH
Lactonase (Pfam-B)
D-Apionate
Oxido-isomerase
Xyl Isom
CO2
L-Erythrulose
DUF1537
DHA Kinase
Tpi
Rpi
D-Erythrose 4P
RLP-1
RLP-2
Glycolate
H2O
3-PGA

Figure 4

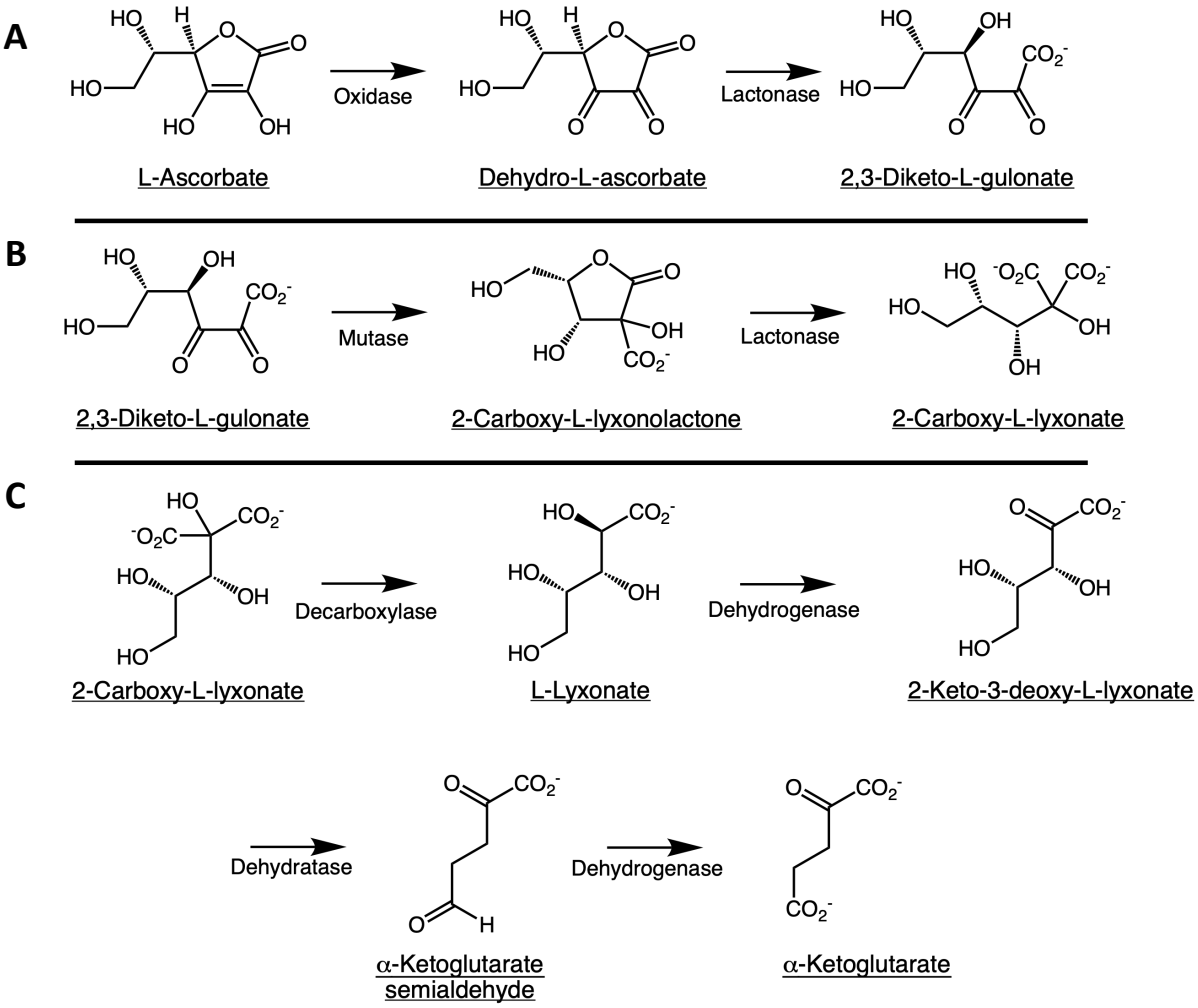

A
Oxidase
Lactonase
L-Ascorbate
Dehydro-L-ascorbate
2,3-Diketo-L-gulonate
B
Mutase
Lactonase
2,3-Diketo-L-gulonate
2-Carboxy-L-lyxonolactone
2-Carboxy-L-lyxonate
C
Decarboxylase
Dehydrogenase
2-Carboxy-L-lyxonate
L-Lyxonate
2-Keto-3-deoxy-L-lyxonate
Dehydratase
Dehydrogenase
α-Ketoglutarate semialdehyde
α-Ketoglutarate

Figure 5

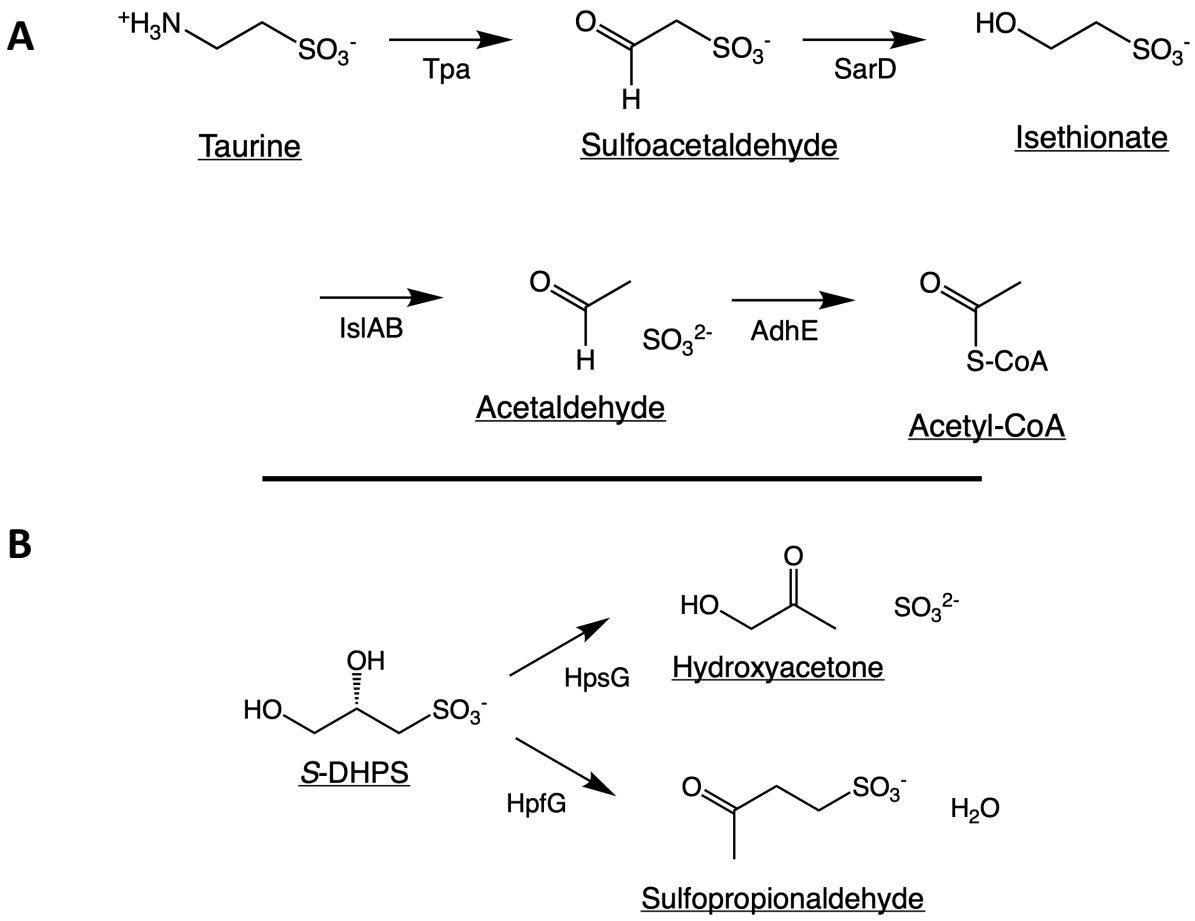

A
+H3N
SO3-
Tpa
O
SO3-
H
SarD
HO
SO3-
Taurine
Sulfoacetaldehyde
Isethionate
IslAB
O
H
SO3 2-
AdhE
O
S-CoA
Acetaldehyde
Acetyl-CoA
B
OH
HO
SO3-
S-DHPS
HpsG
O
HO
SO3 2-
Hydroxyacetone
HpfG
O
SO3-
H2O
Sulfopropionaldehyde

Figure 6

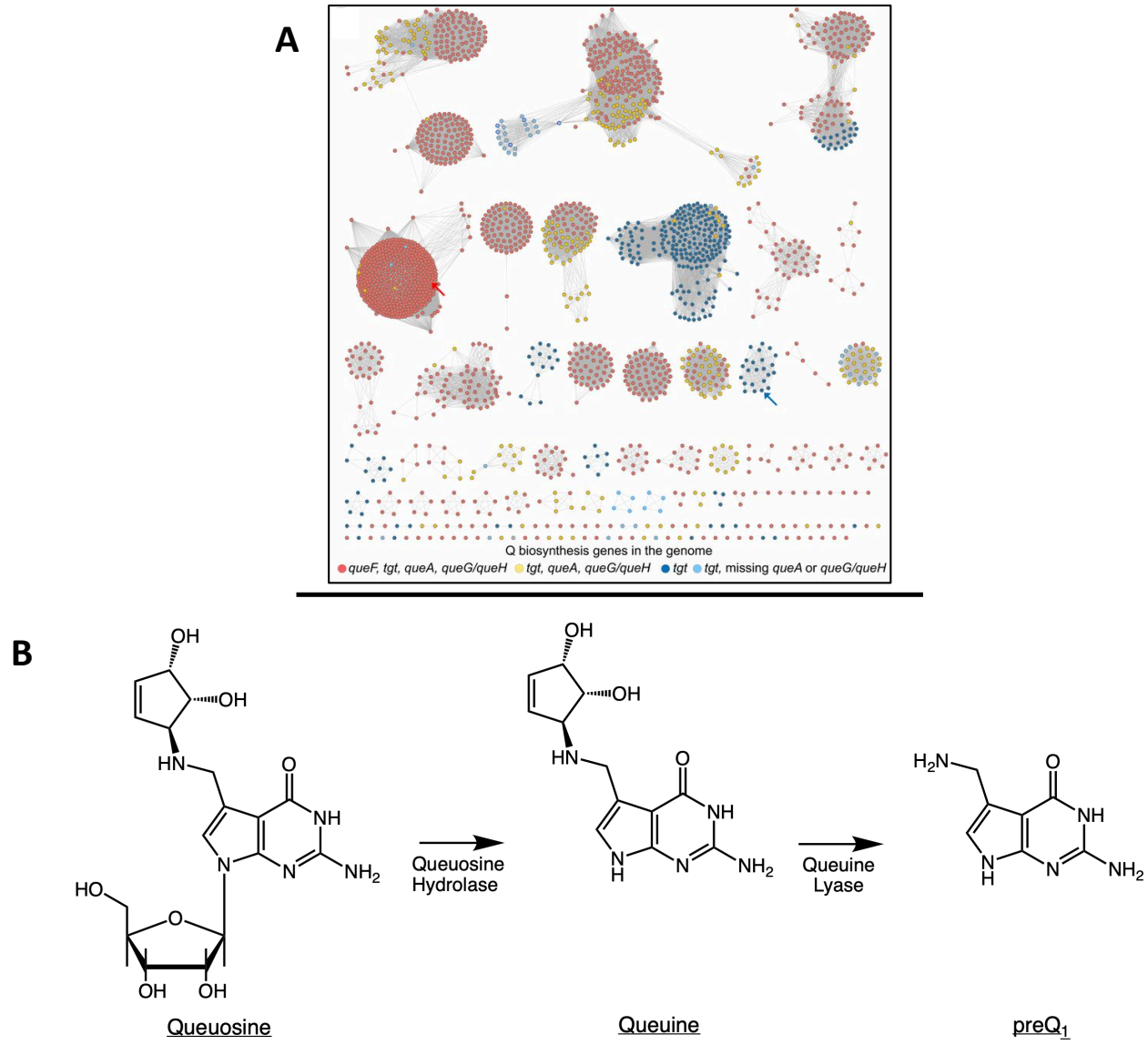
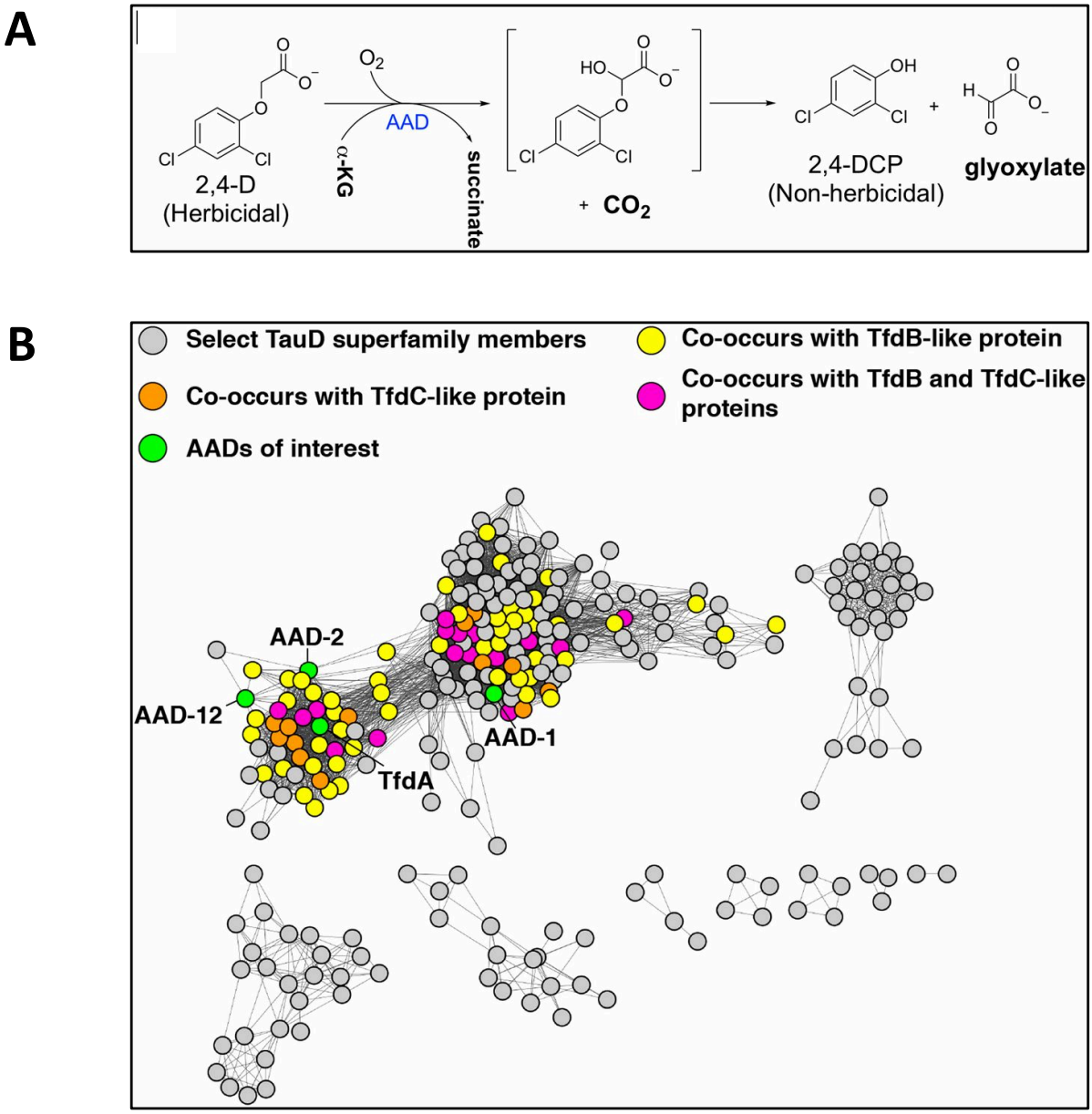

A
Q biosynthesis genes in the genome
queF, tgt, queA, queG/queH
tgt, queA, queG/queH
tgt
tgt, missing queA or queG/queH
B
Queuosine
Queuosine Hydrolase
Queuine
Queuine Lyase
preQ1

Figure 7


A
2,4-D
(Herbicidal)
O2
AAD
α-KG
succinate
+ CO2
2,4-DCP
(Non-herbicidal)
glyoxylate
B
Select TauD superfamily members
Co-occurs with TfdB-like protein
Co-occurs with TfdC-like protein
Co-occurs with TfdB and TfdC-like proteins
AADs of interest
AAD-2
AAD-12
AAD-1
TfdA