Biochemistry

Perspective

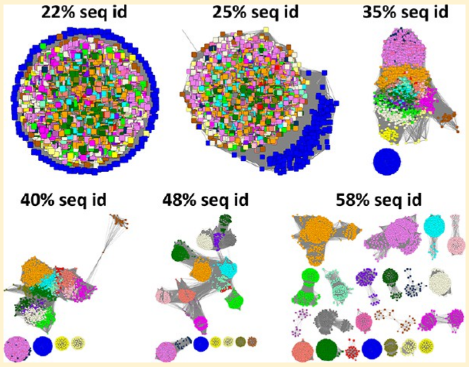
pubs.acs.org/biochemistry

Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence−Function Space and Genome Context to Discover Novel Functions

John A. Gerlt*

Departments of Biochemistry and Chemistry, Institute for Genomic Biology, University of Illinois, Urbana-Champaign Urbana, Illinois 61801, United States

ABSTRACT: The exponentially increasing number of protein and nucleic acid sequences provides opportunities to discover novel enzymes, metabolic pathways, and metabolites/natural products, thereby adding to our knowledge of biochemistry and biology. The challenge has evolved from generating sequence information to mining the databases to integrating and leveraging the available information, i.e., the availability of "genomic enzymology" web tools. Web tools that allow identification of biosynthetic gene clusters are widely used by the natural products/synthetic biology community, thereby facilitating the discovery of novel natural products and the enzymes responsible for their biosynthesis. However, many novel enzymes with interesting mechanisms participate in uncharacterized small-molecule metabolic pathways; their discovery and functional characterization also can be accomplished by leveraging information in protein and nucleic acid databases. This Perspective focuses on two genomic enzymology web tools that assist the discovery novel metabolic pathways: (1) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) for generating sequence similarity networks to visualize and analyze sequence−function space in protein families and (2) Enzyme Function Initiative-Genome Neighborhood Tool (EFI-GNT) for generating genome neighborhood networks to visualize and analyze the genome context in microbial and fungal genomes. Both tools have been adapted to other applications to facilitate target selection for enzyme discovery and functional characterization. As the natural products community has demonstrated, the enzymology community needs to embrace the essential role of web tools that allow the protein and genome sequence databases to be leveraged for novel insights into enzymological problems.



In 2001 Patricia Babbitt and I discussed nature's strategies for divergent evolution of new enzymatic functions from a common progenitor to yield mechanistically diverse enzyme superfamilies (conserved active site architectures that catalyze reactions with shared partial reactions, intermediates, or transition states) and functionally diverse suprafamilies (conserved active site architectures that catalyze mechanistically distinct reactions).[1] When our review was published, only a few superfamilies/suprafamilies had been recognized, including the enolase, amidohydrolase, thiyl radical, enoyl-CoA hydratase (crotonase), vicinal-oxygen-chelate superfamilies, and the orotidine 5′-monophosphate (OMP) decarboxylase suprafamily, not surprising because the UniProt database then contained only 571 804 protein sequences (July 2001) (http://www.uniprot.org/; see Table 1 for a summary of abbreviations). Despite, in retrospect, a meager number of sequences, we concluded that enzymologists were positioned to expand their interests beyond studies of single enzymes to encompass entire enzyme families. We proposed that sequenced genomes (1) provided a rapidly expanding source of new proteins for investigation and (2) allowed genomic context to be used to infer novel enzymatic functions and, therefore, better understand the evolution of functional diversity in enzyme superfamilies. We suggested the term *genomic enzymology* to describe the expansive strategy of using protein families and genome context to focus studies of enzyme mechanisms, discover new functions, and more accurately describe the evolution of enzyme function in molecular terms (sequence and structure). However, we did not propose how the protein and genome sequence databases could be leveraged and used by the experimental community.

Sixteen years later, the UniProt database contains 88 588 026 nonredundant sequences (Figure 1; Release 2017_07); the number of sequences is increasing at the rate of 2.4% per month (doubling time 2.5 years), largely the result of microbial genome projects. The challenge is to devise "user friendly" methods to interrogate the massive amount of data so that hypotheses can be generated that direct experimental determination of *in vitro* activities and *in vivo* metabolic functions of uncharacterized enzymes. For example, 379 mechanistically diverse superfamilies and functionally diverse suprafamilies have been described;[2] additional superfamilies and suprafamilies must be present in (1) genomic "dark matter" that has not been curated by databases such as Pfam and (2) the genomes of phylogenetically diverse bacterial species that have not yet been systematically

Received: June 29, 2017
Revised: August 5, 2017
Published: August 22, 2017

ACS Publications

DOI: 10.1021/acs.biochem.7b00614

Table 1. List of Abbreviations

ABC	ATP-binding cassette
AGeNNT	Automatically Generates refined Neighborhood NeTworks
antiSMASH	Antibiotics & Secondary Metabolite Analysis SHell
BGC	biosynthetic gene cluster
BLAST	Basic Local Alignment Search Tool
DSF	differential scanning fluorimetry
DUF	domain of unknown function
EFI	Enzyme Function Initiative
EFI-EST	EFI-Enzyme Similarity Tool
EFI-GNT	EFI-Genome Neighborhood Tool
ENA	European Nucleotide Archive
GNN	Genome Neighborhood Network
GRE	glycyl radical enzyme
InterPro	Integrated Protein Database
JGI-IMG/M	Joint Genome Institute-Integrated Microbial Genomes/ Metagenomes
MSA	multiple sequence alignment
NCBI	National Center for Bioinformatics Information
NRPS	nonribosomal peptide synthase
OMP	orotidine 5′-monophosphate
orf	open reading frame
P5C	Δ^1-pyrroline-5-carboxylate
Pfam	Protein Family Database
PKS	polyketide synthase
PN	proteome network
PRISM	PRediction Informatics for Secondary Metabolomes
RLP	RuBisCO-like protein
RODEO	rapid ORF description and evaluation online
RuBisCO	ribulose bisphosphate carboxylase/oxygenase
SBP	solute binding protein
SFLD	Structure−Function Linkage Database
ShortBRED	"Short, Better Representative Extract Data Set"
SSN	Sequence Similarity Network
TCT	tricarboxylate transport
TRAP	tripartite ATP-independent periplasmic transporter
TRN	Taxonomic Rank Network
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase

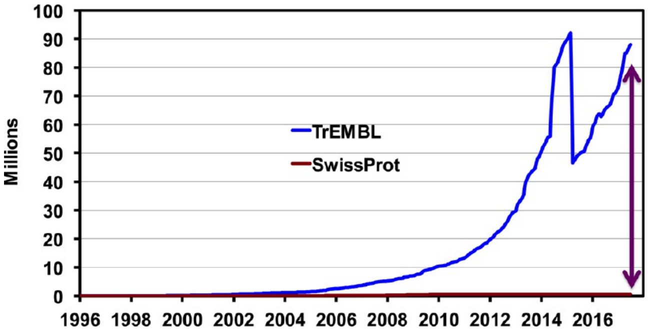


Figure 1. Growth of the UniProt protein sequence database (Release 2017_07). The blue line represents the EMBL/TrEMBL sequences with automated annotations; the red line represents the EMBL/SwissProt with manually curated annotations. Currently, the doubling time is ∼2.5 years. The number of sequences decreased by ∼50% in April 2015 when UniProt identified reference proteomes for closely related species and archived the redundant proteomes.

sequenced.[3] This large, and growing for the foreseeable future, set of superfamilies includes members that catalyze novel reactions in novel pathways, a boon to enzymologists.

Approximately 50% of the proteins in the databases have incorrect, uncertain, or unknown functional annotations.[4] The UniProt Knowledgebase (UniProtKB) is composed of two sections, UniProtKB/SwissProt and UniProtKB/TrEMBL. The annotations in UniProtKB/SwissProt are manually curated; the functional annotations in UniProtKB/TrEMBL are computationally assigned based on the function of the "closest" homologue. In the most recent UniProt release (2017_07), only 0.63% of the sequences are in the UniProtKB/SwissProt section (Figure 1); this fraction continues to decrease because the total number of sequences added in each release greatly exceeds the number of new sequences with SwissProt-curated, experimentally verified annotations. In principle, curated annotations might be extended to orthologues; however, the sequence boundaries between functions are unknown, so homology-based approaches for functional assignment are risky. Therefore, incorrect, uncertain, or unknown annotations will continue to propagate, compromising their utility to allow the discovery of new enzymatic functions, metabolic pathways, metabolites, and biology.

Khosla recently summarized this challenge:[5] "Although enzymology will remain a predominantly experimental science for the foreseeable future, one cannot avoid a sense of helplessness when one considers the huge (and growing) deficit in functionally annotated sequences. By now, there are approximately 100 million nonredundant protein sequence entries in GenBank, but a reliably curated protein database such as SwissProt contains fewer than 1 million entries. This is a quintessential 'big data' problem, where the rate at which data is generated continues to outpace the rate at which it is curated. It is unlikely that more resource-intensive curation alone can solve the problem. As the proverb says, this may be a situation where the most desirable approach will involve user-friendly tools that teach a novice how to fish instead of serving fish. Such tools could ideally capture the essence of an enzymologist's judgment in layers of increasing sophistication, depending on the user's actual needs."

This Perspective describes "genomic enzymology" web tools that initially were developed by the Enzyme Function Initiative (EFI)[6] and provides examples of their applications.

Web Tools for Natural Product Discovery. In parallel with the development of genomic enzymology, the natural products community discovered that genes encoding biosynthetic pathways for natural products often are organized in "biosynthetic gene clusters" (BGCs).[7−9] Given the structural complexity of natural products and the need to identify the enzymes that assemble their backbones, e.g., terpene synthases, nonribosomal peptide synthases (NRPSs), and polyketide synthases (PKSs), as well as the enzymes that catalyze "tailoring" reactions, e.g., glycosylases, methylases, and redox enzymes, the genomic colocalization of the biosynthetic genes facilitates pathway discovery and experimental characterization. Although the type of scaffold may be apparent from the annotations in the BGCs, the structure of the natural product is not trivial to predict. Indeed, many enzymes (backbone-forming and tailoring) are novel members of diverse enzyme superfamilies. Nonetheless, the discovery of a BGC facilitates enzyme identification so that they can be experimentally tested for sequential activities in the biosynthetic pathway.

The number of natural products is estimated to be extremely large;[10,11] therefore, identification of BGCs is an attractive strategy for their discovery. In the past several years, bioinformatic tools have been developed for discovering BGCs

DOI: 10.1021/acs.biochem.7b00614

in sequenced genomes,[12,13] including antiSMASH (Antibiotics & Secondary Metabolite Analysis SHell[14]), PRISM (PRediction Informatics for Secondary Metabolomes[15]), and RODEO (Rapid ORF Description and Evaluation Online[16]). These tools are widely used by the natural products/synthetic biology community, e.g., more than 300 000 jobs have been processed by the antiSMASH server (https://antismash.secondarymetabolites.org/). Although these tools enable the discovery of BGCs, the annotations of the uncharacterized enzymes in the BGCs are limited to their membership in protein families, an overview that often is insufficient to restrict substrate specificities and/or reaction identities/mechanisms. Therefore, many of the challenges in BGC characterization are the same as those encountered by enzymologists focused on small-molecule metabolic pathways (*vide infra*).

What Should Genomic Enzymology Tools Provide? Genomic enzymology focuses on the discovery of function in the context of entire enzyme families: this approach allows recognition of sequence and structure attributes that are conserved for specific functions. Babbitt developed the Structure−Function Linkage Database (SFLD; http://sfld.rbvi.ucsf.edu/) to generate and disseminate sequence−structure relationships that associate specific functional properties with specific sequence and structure motifs in functionally diverse enzyme superfamilies.[17] As an early example of the use of genomic enzymology to obtain mechanistic insights, the recognition that (1) the reactions catalyzed by mandelate racemase and muconate lactonizing enzyme in the enolase superfamily require stabilization of an enolate anion intermediate and (2) their sequences have conserved motifs for binding an active site Mg^{2+} defined the catalytic strategy for the superfamily.[1,18,19] The functional diversity in the superfamily, including dehydration, deamination, cycloisomerization, racemization, and epimerization of carboxylate-anion substrates, could be explained by divergent evolution selecting (1) acid/base catalysts for both generating the enolate anion intermediate and directing it to products and (2) specificity determinants for binding different substrates in productive geometries relative to the acid/base catalysts.[20,21] This same strategy for evolution of new enzymatic functions applies to many mechanistically diverse superfamilies.[2]

The challenges for genomic enzymology are developing and applying *large-scale* methods for (1) grouping members of mechanistically diverse superfamilies and functionally diverse suprafamilies in isofunctional families, e.g., identifying acid/base catalysts and placing restrictions on reaction mechanisms and substrate specificities and (2) analyzing the genome contexts for the members of isofunctional families so that their roles in metabolic pathways can be deduced. e.g., predicting substrates, intermediates, and products.

Sequence Similarity Networks (SSNs). Evolutionary biologists typically use phylogenetics-based approaches to distinguish orthologues from paralogues.[22,23] Phylogenetic trees are constructed from multiple sequence alignments (MSAs); however, MSAs are difficult to generate for large protein families.[23] Many superfamilies and suprafamilies are large: >15 K sequences in the glycyl-radical enzyme superfamily, >22 K sequences in the OMP decarboxylase suprafamily, >44 K sequences in the enolase superfamily, >122 K sequences in the enoyl-CoA hydratase (crotonase) superfamily, and >250 K sequences in the radical SAM superfamily. In addition to being difficult to construct, trees for large families also are difficult to interpret because of their complexity.[24] Trees do not provide immediate access to all sequences in a family—representative sequences usually are selected in the construction of the tree. Instead, what is needed is a large-scale approach that allows easy visualization and analyses for all sequences in a family, recognizing that it must be "user friendly", i.e., intuitive and fast.

Atkinson and Babbitt introduced sequence similarity networks (SSNs) to enable large-scale analyses of sequence−function relationships in protein families.[25] An SSN displays pairwise relationships obtained from an all-by-all sequence comparison, e.g., BLAST. Although the use of BLAST can be criticized because it provides a measure of overall sequence similarity and, therefore, may be insensitive to different domain architectures important in determining molecular function, it is (1) fast, a requirement for routine all-by-all comparisons of the sequences of members of increasingly large protein families (each sequence must be compared with every other sequence so the time required increases with the square of the number of sequences), and (2) familiar to experimentalists. An SSN contains "nodes" for sequences; "edges" that quantitate sequence similarity (pairwise sequence identity) connect nodes that share sequence similarity that exceeds a user-specified level (Figure 2). As the sequence similarity required to connect nodes with edges is increased, the nodes segregate into clusters; the goal is to select a level of sequence similarity that segregates the nodes/members of the family into isofunctional clusters (Figure 3).

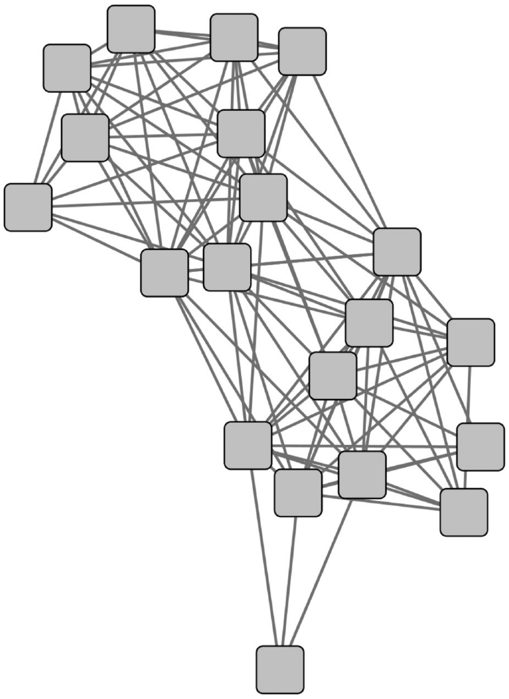


Figure 2. A sequence similarity network (SSN) showing the protein sequence nodes and pairwise sequence similarity edges.

SSNs contain "node attributes", including functional and phylogenetic information associated with each sequence/node, that assist the user in analyzing sequence−function relationships, including choosing sequence similarity thresholds for drawing edges and segregating the families into isofunctional clusters. Atkinson and Babbitt compared SSNs with phylogenetic trees and concluded "the most valuable feature of SSNs is not the optimal or most accurate display of sequence similarity, but rather the flexible visualization of many alternate protein attributes for all or nearly all sequences in a superfamily".[25]

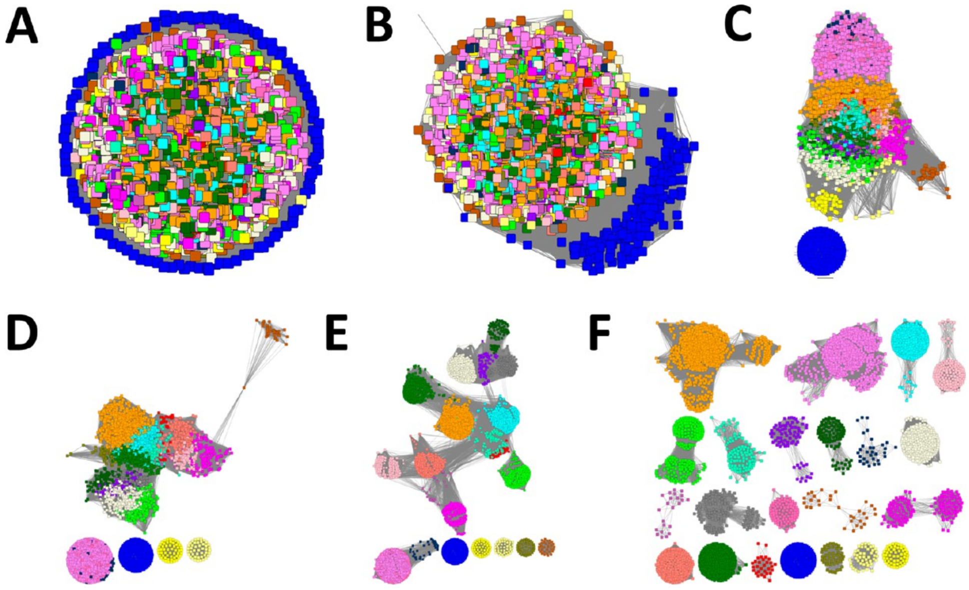


Figure 3. SSNs for sequences from the proline racemase family (Pfam family PF05544). (A) Alignment score ≥15, ≥22% pairwise sequence identity. (B) Alignment score ≥20, ≥25% pairwise sequence identity. (C) Alignment score ≥50, ≥35% sequence identity. (D) Alignment score ≥70, ≥40% sequence identity. (E) Alignment score ≥90, ≥48% sequence identity. (F) Alignment score ≥110, ≥58% sequence identity. The colors in panel F are used to color the nodes in panels A−E.

SSNs are viewed using Cytoscape (http://cytoscape.org/), "an open source platform for visualizing complex networks and integrating these with attribute data".[26] Although Cytoscape has a steep "learning curve", it provides Control Panels to select nodes based on the node attributes and to filter and color the networks to enable visual analyses. With node attributes and the Control Panels, SSNs viewed with Cytoscape satisfy Khosla's vision that genomic enzymology tools "could ideally capture the essence of an enzymologist's judgment in layers of increasing sophistication, depending on the user's actual needs".[5]

The SFLD provides SSNs for a several functionally diverse superfamilies with manually curated (labor intensive and expensive) annotations/node attributes;[17] these SSNs serve as "gold standards" for functional annotation in both the bioinformatics and enzymology communities.[27] However, with the large number of superfamilies/suprafamilies (*vide infra*) and families that provide additional metabolic enzymes, e.g., dehydrogenases, kinases, and aldolases, community-initiated generation of SSNs is necessary. The SFLD does not provide this capability; Pythoscape was developed by the SFLD for generating large SSNs, but it is not "user friendly" for most experimentalists because it requires access to a computer cluster and programming expertise.[28]

In principle, the construction of SSNs is "simple", i.e., connecting sequences with edges that quantitate similarity. However, most experimentalists would be hard-pressed to develop their own programs for generating SSNs. And, other web tools that construct SSNs, e.g., Pclust[29] and CLANS,[30] use a limited number of sequences and/or node attributes.

The EFI developed a web tool, the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST; http://efi.igb.illinois.edu/efi-est/),[31] to generate SSNs for large protein families. To date, >1600 unique users have submitted jobs to EFI-EST, and >50 publications have appeared that reference the use of EFI-EST.[13,14,32−78] EFI-EST uses sequences and node attribute information from UniProt: in contrast to the NCBI database, annotations in the UniProt database can be changed with data provided by any member of the community, allowing important corrections and additions that diminish propagation of annotation errors.

EFI-EST now provides four options for selecting sequences to be included in the SSN: Option A, a single user-supplied sequence is used to collect homologues with BLAST from the UniProt database (maximum 10 000 sequences); Option B, the user specifies one or more UniProt and/or InterPro families [currently limited to ≤255,000 sequences to allow the SSN for the radical SAM superfamily (Pfam family PF04055) to be generated]; Option C (enhanced in the most recent update), the user provides a FASTA file of sequences and selects whether accession IDs in the headers are used to retrieve node attributes from UniProt; and Option D (new in the most recent update), the user provides a list of UniProt and/or NCBI accession IDs. After the all-by-all comparison using BLAST, the user selects an "alignment score" based on pairwise percent identity to filter the edges (the threshold for drawing edges to connect nodes). The user then downloads the SSN for analysis with Cytoscape.

EFI-EST now provides a "Color SSN Utility" to facilitate analyses of SSNs by (1) coloring each cluster in an input SSN with a unique color, (2) providing a file with color information that allows the user to color SSNs of the same sequences generated with lower similarity (pairwise identity) to track segregation of clusters (e.g., Figure 3), and (3) FASTA files for the sequences in each cluster to facilitate the generation of MSAs.

Applications of SSNs. The EFI used SSNs from the SFLD to characterize sequence−function space in targeted functionally diverse superfamilies (amidohydrolase,[79−85] enolase,[19,86−92] glutathione *S*-transferase,[93] haloalkanoate dehalogenase,[94] and isoprenoid synthase[95,96]) and select targets for functional discovery. Then, when EFI-EST became available, both the EFI and community began to use SSNs to characterize sequence−function space in a wide range of proteins families.

SSNs generated by the community using EFI-EST[13,14,32−78] have been used to identify and describe potential isofunctional families within enzyme families, e.g., clusters with different (but

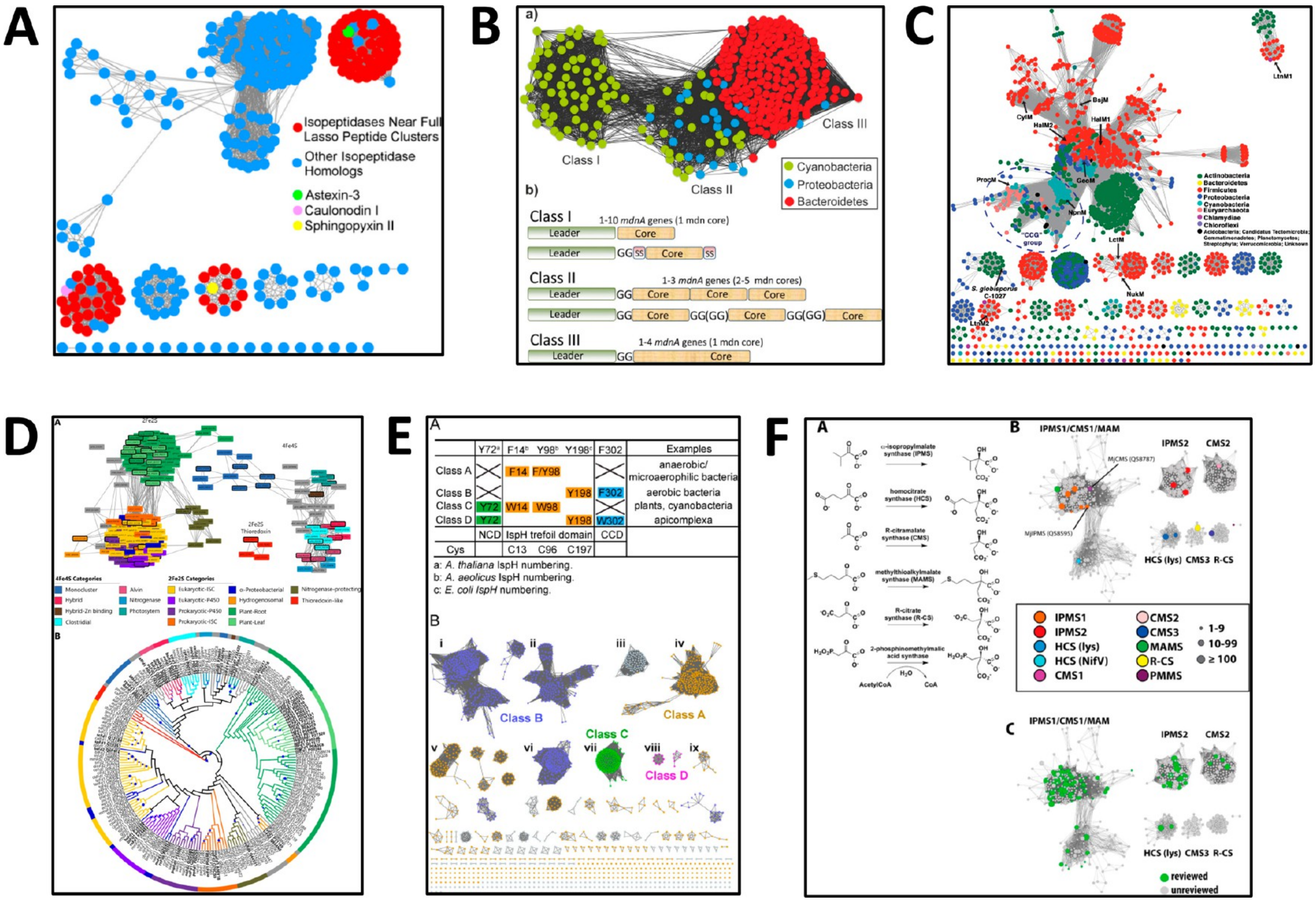


Figure 4. Examples of SSNs generated with EFI-EST that were included in recent publications. (A) SSN for isopeptidases involved in lasso peptide synthesis.[43] (B) SSN of precursor peptides for microviridin synthesis.[60] (C) SSN of LanMs in lantibiotic synthesis.[76] (D) SSN for ferredoxins compared with a phylogenetic tree.[40] (E) SSN for IspH in isoprenoid biosynthesis.[56] (F) SSNs for members of the DRE-TIM metallolyase superfamily.[52] Figures reproduced with permission from refs 40, 43, 52, 56, 60, and 76.

unknown) substrate specificities, thereby providing an overview of sequence−function space in specificity diverse superfamilies (different substrates but same type of overall reaction) and functionally diverse superfamilies (different substrates and different reaction mechanisms, although a partial reaction may be conserved). SSNs also provide the ability to survey the members of a protein family for different domain architectures that may suggest different functional contexts, i.e., fusion proteins in different pathways. And, the pathway for cluster segregation as sequence similarity increases (Figure 3) may suggest functional linkages between clusters. Several community-generated SSNs from the recent literature that illustrate their use are shown in Figure 4; readers are referred to the publications for detailed descriptions.[13,14,32−78]

Genome Neighborhood Networks (GNNs). With the potential to segregate protein families into isofunctional clusters using SSNs, the second genomic enzymology challenge is to place these clusters in a functional context, e.g., identify the small-molecule metabolic pathways in which uncharacterized enzymes participate. In eubacteria, archaea, and fungi, the enzymes in a metabolic pathway often are encoded by a gene cluster or operon (just as the biosynthetic pathways for natural products are encoded by BGCs). Therefore, the proteins encoded by the genes proximal to those that encode members of an isofunctional cluster (orthologues) may allow the number and types of reactions in the metabolic pathway to be determined if these are conserved by the members of the cluster.

Genome neighborhoods for homologues can be examined using web resources such as JGI-IMG/M (https://img.jgi.doe.gov/cgi-bin/m/main.cgi); however, complete pathways are not always encoded by a single genome neighborhood. Large-scale mining of genome neighborhoods for all orthologues in an SSN cluster has the advantage that operon/gene cluster organization may not be preserved across phylogenetic species; i.e., the sequences in an isofunctional SSN cluster may have diverse genome neighborhoods and pathway neighbors, but the ability to survey all of the neighborhoods provides the potential to identify all of the functionally linked genes/enzymes that can be assembled into a metabolic pathway.

In 2014, the EFI described a genome neighborhood analysis that was applied to the proline racemase family (Pfam family PF05544) using an all-by-all comparison (with BLAST) of the neighbors to generate a network (the genome neighborhood network, GNN);[97] the neighbors were segregated into protein families using an e-value >20 for the edges in the SSN. By assigning unique colors to the clusters in the SSN (Figure 5A) and coloring the neighbors in the GNN with the same color, the neighbors for the sequences in each cluster were identified (Figure 5B). Then, candidates for functionally linked enzymes were recognized and potential pathways were predicted. This analysis allowed *in vitro* enzymatic activities and *in vivo* metabolic

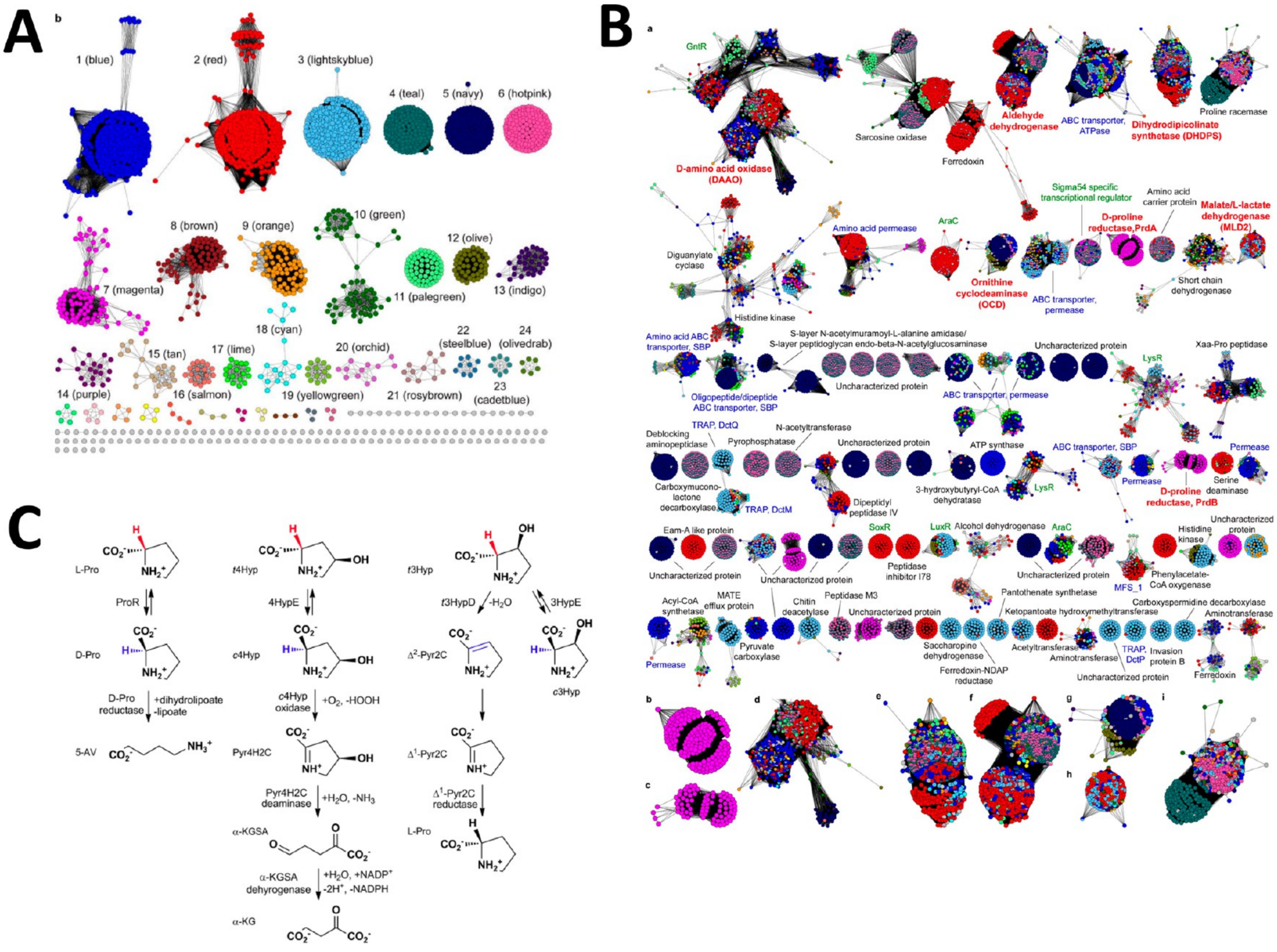


Figure 5. (A) A colored SSN for the proline racemase family (PF05544; InterPro Release 43.0). (B) The GNN generated by an all-by-all BLAST of the genome neighbors. (C) Three pathways catalyzed by members of the proline racemase family. The nodes in the GNN (panel B) are colored using the color clusters in the SSN (Panel A). Figures reproduced with permission from ref 97.

functions (the three pathways shown in Figure 5C) to be assigned to 85% of the sequences in the family [2333 sequences in InterPro Release 43.0 (July 2013)].

The EFI subsequently developed the Enzyme Function Initiative-Genome Neighborhood Tool (EFI-GNT; http://efi.igb.illinois.edu/efi-gnt/) to provide a "user friendly" interface for generating GNNs to facilitate the identification of pathway/metabolic context for isofunctional clusters in SSNs. Although EFI-GNT has not yet been "officially" announced with a detailed publication (a manuscript describing the updated version of EFI-EST and EFI-GNT is in preparation for publication later this year), >250 unique users have accessed the web tool that is available for community use.

An SSN generated by EFI-EST is the input for EFI-GNT [Figure 6A; 6419 sequences in the proline racemase family in InterPro Release 63.0 (May 2017)]. EFI-GNT assigns a unique color (from a palette of 1513 colors) to each cluster (Figure 6B). It then interrogates the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) database for the neighbors of each sequence in each cluster in the input SSN (for eubacteria, archaea, and fungi), and the neighbors are associated with their Pfam families. The co-occurrence frequencies of the queries in the SSN cluster with the neighbors as well as the absolute values of the distances in open reading frames (orfs) between the queries and neighbors are calculated. Functionally linked genes encoding a pathway are expected to have (1) large query-neighbor co-occurrence frequencies (diminished if operon/gene cluster organization is phylogenetically diverse) and (2) short distances between the queries and neighbors.

EFI-GNT provides GNNs in two formats. In one format (Figure 6C,D), a cluster is present for each SSN cluster: the hub-node represents the sequences in the SSN cluster (colored with a unique color so that it can be easily identified in a colored version of the input SSN that is generated), and the spoke-nodes represent the neighbor Pfam families; this format allows the user to identify the pathway enzymes. In the second format, a cluster is present for each neighbor Pfam family: the hub-node represents the Pfam family, and the spoke nodes represent the SNN clusters that identified the neighbors (Figure 6E,F); this format allows the user to assess whether the similarity (edge) threshold used to generate the input SSN was too large (pairwise identity too large) so that orthologues are segregated in multiple clusters, with these identifying the same Pfam family neighbors and pathway.

In both GNN formats, the co-occurrence frequencies of the SSN queries and neighbors are the values of the edges between the hub- and spoke-nodes: if the co-occurrence frequency exceeds a user-specified threshold, the edge and spoke-node are

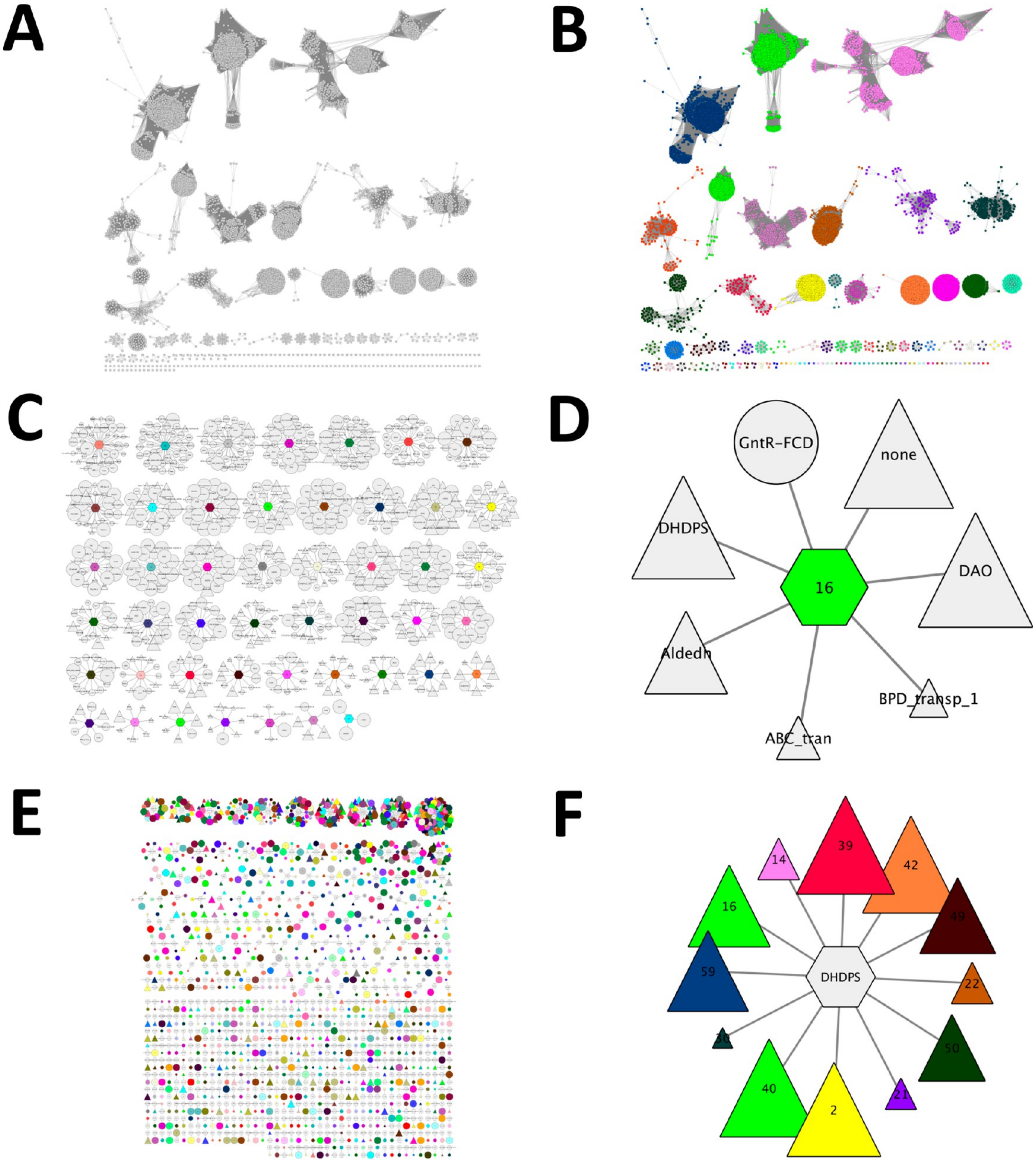


Figure 6. (A) SSN for the proline racemase family (PF05544, InterPro Release 63.0) segregated with an alignment score of ≥110 (≥58% pairwise sequence identity). (B) Colored SSN generated by the EFI-GNT web tool. (C, D) GNN with SSN cluster hub-nodes and Pfam family spoke-nodes. (E, F) GNN with Pfam family hub-nodes and SSN cluster spoke-nodes. The GNNs were generated with a ±10 orf genome neighborhood window and a query-neighbor co-occurrence threshold of 20%.

present. From the co-occurrence frequencies, the user can identify neighbors that "always" occur with the query (the same conserved operon/gene cluster) as well as those that are less frequently associated (operon/gene cluster in some species; dispersed genes in other species).

EFI-GNT also provides files with the UniProt IDs for the sequences in each neighbor Pfam family that can be used to identify the neighbors in the SSNs for their families. This mapping (1) assists the selection of alignment score thresholds for segregating the neighbor SSNs into isofunctional clusters/families and (2) provides useful context about possible functional

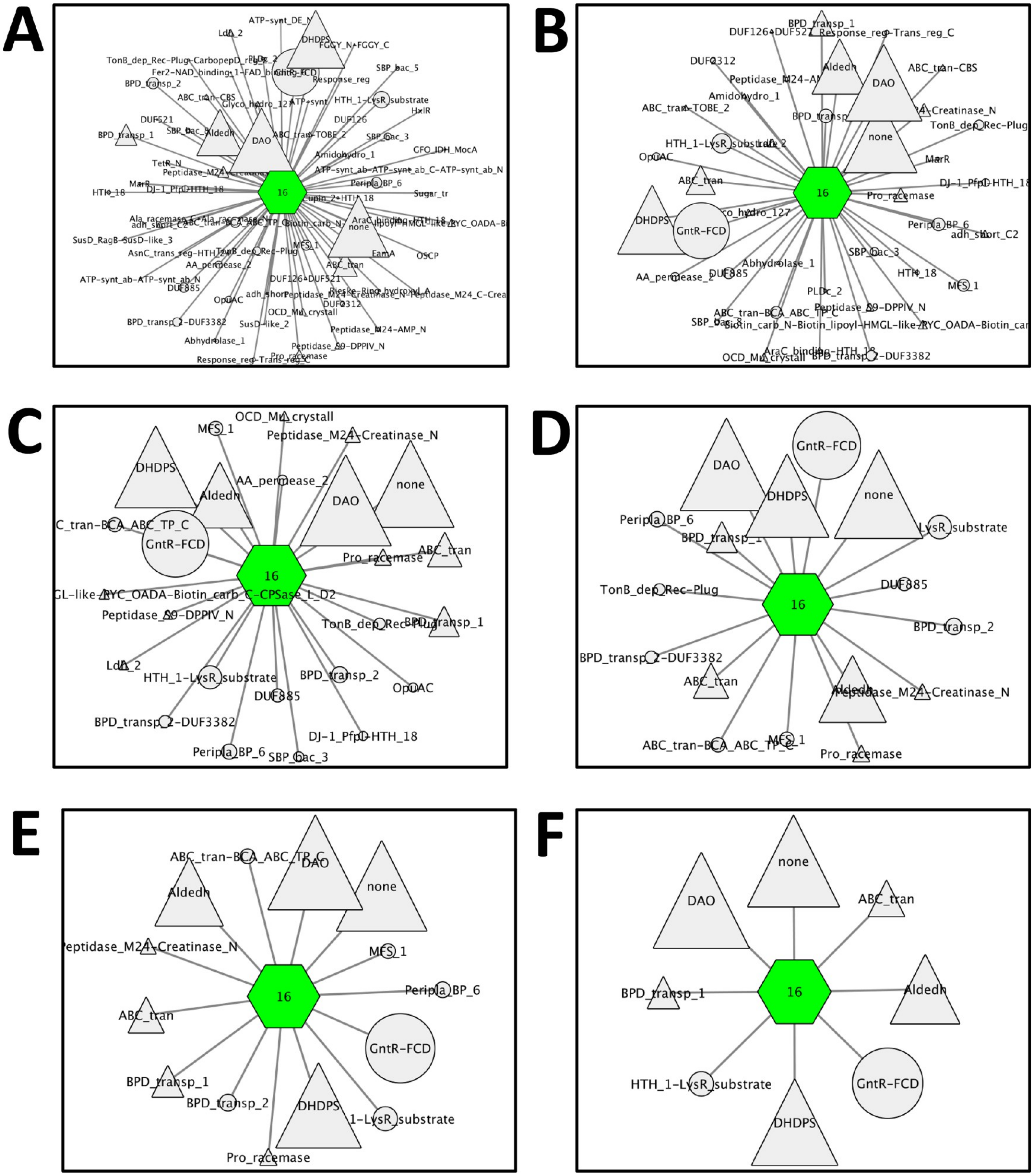


Figure 7. GNN for SSN cluster 16 presented at different query-neighbor co-occurrence frequencies. (A) 3%. (B) 5%. (C) 10%. (D) 12%. (E) 15%. (F) 20%.

(substrate specificity and reaction mechanism) relationships that may be useful in deducing *in vitro* activities and *in vivo* metabolic functions.

Integrated Use of SSNs and GNNs To Discover Metabolic Pathways. The synergistic "power" of the EFI-EST and EFI-GNT web tools for functional annotation of bacterial and fungal enzymes is the ability to (1) segregate protein families into isofunctional clusters in an SSN using EFI-EST (the sequences in a cluster have the same genome context) and (2) use the SSN as the input for EFI-GNT to interrogate and visualize genome neighborhood context for the isofunctional clusters in the GNN. To the best of our knowledge, no other web tools provide this integrated capability.

DOI: 10.1021/acs.biochem.7b00614

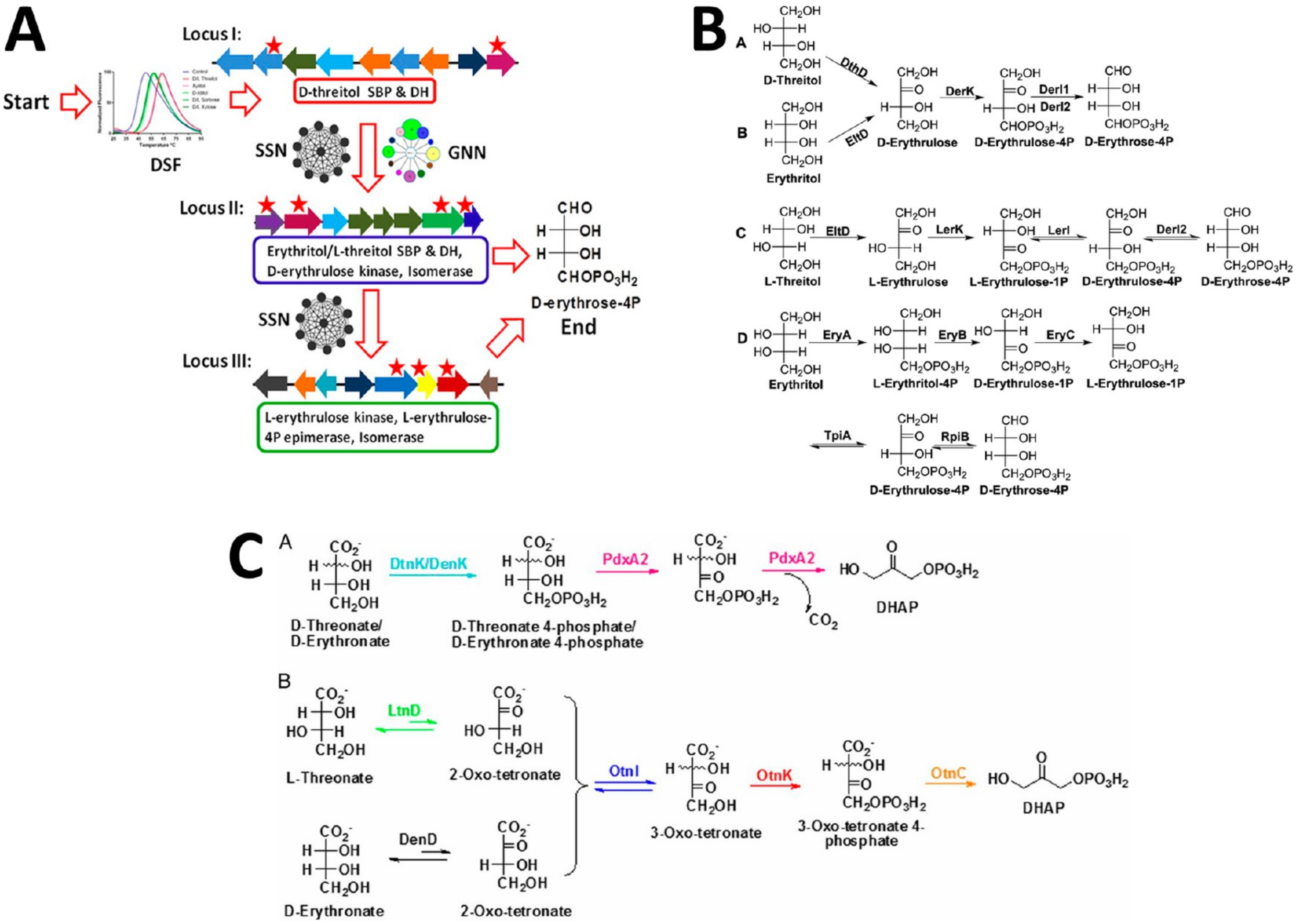


Figure 8. (A) Strategy for discovering catabolic pathways for D-threitol, L-threitol, and erythritol in *M. smegmatis* using differential scanning fluorimetry (DSF) to screen the ligand specificities of SBPs and the integrated used of SSNs and GNNs to discover the pathway enzymes. (B) Catabolic pathways for D-threitol, L-threitol, and erythritol. (C) Catabolic pathways for D-threonate, L-threonate, and D-erythronate in *R. eutropha* H16.[59] Figures in Panel A and B reproduced with permission from ref 34; figure in Panel C reproduced with permission from ref 59.

The GNN format in which the hub-node represents the SSN cluster and the spoke-nodes represent the Pfam families (Figure 6C,D) can be used to identify the enzymes, transcriptional regulators, and transporters in a metabolic pathway. For example, continuing with the proline racemase family (PF05544; SSN in Figure 6A,B), the enzymes in a catabolic pathway for the conversion of *trans*-4-hydroxyproline to *α*-ketoglutarate (middle pathway in Figure 5C) can be identified for cluster 16 in the input SSN (Figure 6D, 792 sequences with genome neighborhoods in the ENA files). In addition to 4-hydroxyproline epimerase (the queries in cluster 16 and the SSN hub-node in the GNN cluster in Figure 6D), the Pfam family spoke-nodes of the GNN cluster identify the three remaining enzymes in the pathway: (1) *cis*-4-hydroxyproline oxidase, a member of the D-amino acid oxidase family ("DAO" in Figure 6D; PF01266, co-occurrence frequency, 0.91, median distance 1.0 orfs); (2) *cis*-4-hydroxyproline imino acid dehydratase/deaminase, a member of the dihydrodipicolinate synthase family ("DHDPS"; PF00701, co-occurrence frequency, 0.82, median distance 2.0 orfs); and (3) *α*-ketoglutarate semialdehyde dehydrogenase, a member of the aldehyde dehydrogenase family ("Aldedh"; PF00171, co-occurrence frequency, 0.66, median distance 2.0 orfs). The curations provided by Pfam provide essential clues for deducing the identities of the reactions catalyzed by the various neighboring enzymes (conserved reaction mechanisms).

The GNN in Figure 6D also includes (1) the ATP-bonding component of an ABC transport system ("ABC_trans", PF00005, co-occurrence frequency, 0.35, median distance 4.0 orfs), (2) an additional membrane component of the ABC transport system ("BPD_transp_1", PF00528, co-occurrence frequency, 0.31, median distance 3.0 orfs), and (3) a bidomain transcriptional regulator ("GntR-FCD", PF00392 and PF07729, co-occurrence frequency, 0.67, median distance 3.0 orfs).

The GNN analysis also recognizes genome neighbors that are not associated with any Pfam family ("none" in Figure 6D; ∼15% of the proteins in UniProt are not associated with a Pfam family). These sequences can contain protein families currently not curated by Pfam; these families can be defined by generating SSNs for these sequences using Option D of EFI-EST.

The GNN in Figure 6D was generated with a minimum co-occurrence frequency of 0.30. At lower co-occurrence frequencies (Figure 7), members of four families of solute binding proteins [SBPs; Peripla_BP_6 (PF13458), SBP_bac_3 (PF00497), Peripl_BP_8 (PF13416), and SBP_bac_5 (PF00496)] for ABC transport systems also are genome proximal to the SSN queries with co-occurrence frequencies of 0.16, 0.11, 0.07, and 0.03, respectively, and median distances of 6.0, 5.0, 2.0, and 6.0 orfs, respectively. Also members of the major facilitator superfamily (MFS_1, PF07690) and an amino acid permease family (AA_permease_2 family, PF13520) are genome proximal to the SSN queries with co-occurrence frequencies of 0.15 and 0.11, respectively, and median distances of 9.0 and 2.0 orfs, respectively. The enzymes in metabolic pathways usually are conserved (orthologues instead of analogues; *vide infra*), but transport systems and transcriptional regulators often are not conserved, so members of multiple families of transporters and regulators may be genome proximal to the queries in the SSN cluster.

Figure 7 illustrates the ability of GNNs to analyze genome neighborhoods as a function of co-occurrence frequency, thereby allowing the identification of pathways that may be encoded by single genome neighborhoods in some species and multiple genome neighborhoods in other species. An example of the utility of this capability is described in the next section.[34]

Use of Transport System SBPs To Anchor Pathway Prediction Using SSNs and GNNs. For uncharacterized pathways, pathway prediction is facilitated by independent information about the substrate for the first enzyme in the pathway. For microbial enzymes in catabolic pathways, such information can be obtained from the identity of the solute for the transporter (or the ligand for a transcriptional regulator). For ABC, TRAP, and TCT transport systems, the solute is conveyed to the membrane components with a soluble extracellular (Gram-positive)/periplasmic (Gram-negative) solute binding protein (SBP); SBPs can be purified on large scale and subjected to ligand screening with differential scanning fluorimetry (DSF)/ThermoFluor using a physical library of small molecules.[98] These ligand specificities anchor the pathway by identifying the substrate for the first enzyme; the Pfam families of the neighbors allow the reactions to be predicted. Experiments, both *in vitro* and *in vivo*, are required to validate the pathway.

Using this strategy, experimentally determined ligands for SBPs and synergistic use of SSNs and GNNs to identify pathway components, the EFI identified several novel catabolic pathways. A particularly informative example is the discovery of catabolic pathways for the three tetritols, D-threitol, L-threitol, and erythritol, in *Mycobacterium smegmatis*.[34] Ligand screening identified one SBP for an ABC transporter that bound D-threitol; a genome-proximal dehydrogenase catalyzed its oxidation; however, other catabolic enzymes were encoded elsewhere in the genome (Figure 8A). These "missing" enzymes were discovered by first constructing the SSN for the D-threitol dehydrogenase and then the GNN for the cluster containing the dehydrogenase—this identified a D-erythrulose kinase that was encoded by a gene cluster distal to the one containing the SBP and D-threitol dehydrogenase in *M. smegmatis* (but not other species that encode the pathway). The SSN for the kinase family was then constructed, and the cluster containing the D-erythrulose kinase was used to construct the GNN; this identified a second gene cluster distal to both the one containing the SBP and D-threitol dehydrogenase and the one containing the D-erythrulose kinase that contained isomerases to complete the D-threitol pathway. Investigation of other genes in both distal clusters allowed identification of the remaining enzymes in the pathway for D-threitol catabolism as well as the enzymes in the pathways for L-threitol and erythritol catabolism (Figure 8B). The ligand specificity of a single SBP was sufficient to identify enzymes for three catabolic pathways encoded by three distal gene clusters.

The EFI also used this strategy to assign functions to members of Domain of Unknown Function 1537 (DUF 1537; approximately 20% of the 16 712 Pfam families in Release 31.0 are families of DUFs or proteins of unknown function).[59] Using the specificities for four SBPs for TRAP transport systems for four-carbon acid sugars, including D-erythronate and L-erythronate, SSNs and GNNs were used to identify two genome neighborhoods in *Ralstonia eutropha* H16 that encode enzymes in catabolic pathways for D-threonate, L-threonate, and D-erythronate (Figure 8C). Members of the DUF1537 family (Pfam families PF07005 and PF17402) were determined to be kinases for four-carbon acid sugars, identifying a previously uncharacterized family of kinases. In addition, members of the PdxA2 family (PF04166) were determined to be oxidative decarboxylases that generate dihydroxyacetone phosphate (DHAP) and CO_2.

In unpublished work, the specificities of three ABC SBPs for D-apiose, a branched chain pentose found in plant cell walls, and the iterative use of SSNs and GNNs have been used to discover five catabolic pathways for D-apiose, a branched aldose, two of which are found in species in the human gut microbiome (humans ingest plant cell walls; species of Bacteroides can degrade the rhamnogalacturonan-II component that contains D-apiose to release D-apiose that can be catabolized[99]). Two pathways include novel RuBisCO-like proteins (RLPs) from the RuBisCO superfamily, one catalyzes a β-ketoacid decarboxylation and the second catalyzes a "transcarboxylation" in which the substrate is decarboxylated (β-ketoacid decarboxylation), with the sequestered CO_2 used to carboxylate the enediolate intermediate on the adjacent carbon, and the resulting isomeric β-ketoacid undergoes hydrolysis as in the canonical RuBisCO reaction. The experimentally determined specificity of three SBPs anchored discovery of five pathways by identifying the substrates; the iterative use of SSNs and GNNs identified the enzymes.

Comments. The success of the integrated application of SSNs and GNNs to discover metabolic pathways is limited by the proximities of the genes encoding the pathway components, so this analysis may not be successful for all functional assignment problems. However, the large-scale nature of the analyses provides the potential to determine whether colocalization of genes is due to limited genetic drift among similar genomes or pathway conservation among phylogenetically diverse genomes; it also allows identification of low co-occurrence frequency but significant clustering of the genes encoding multiple pathway components that would be tedious to discover by examination of large numbers of individual genome neighborhoods.[34]

Also, SSNs provide the ability to segregate members of mechanistically diverse superfamilies and functionally diverse suprafamilies into isofunctional clusters (families). For enzymes an important test of isofunctionality is that the GNN generated for an SSN cluster identifies the components of a single pathway. The iterative use of SSNs and GNNs not only provides a test of isofunctionality but also a method for determining the minimum SSN alignment score required to achieve isofunctionality. If the GNN for an SSN cluster identifies "too many" components for a single pathway, further segregation of the cluster with a larger alignment score into "daughter" clusters may allow the resolution of the pathways. The reader should recognize that achieving isofunctional clusters in an SSN may not be straightforward, e.g., even within the same superfamily different alignment scores may be required to achieve isofunctional clusters. However, the integration of SSNs and GNNs using EFI-EST and EFI-GNT provides a powerful strategy for assessing and achieving isofunctional clusters.

Chemically Guided Functional Profiling: Building on EFI-EST. With ~50% of the proteins in the sequence databases having incorrect, uncertain, or unknown functions, devising a

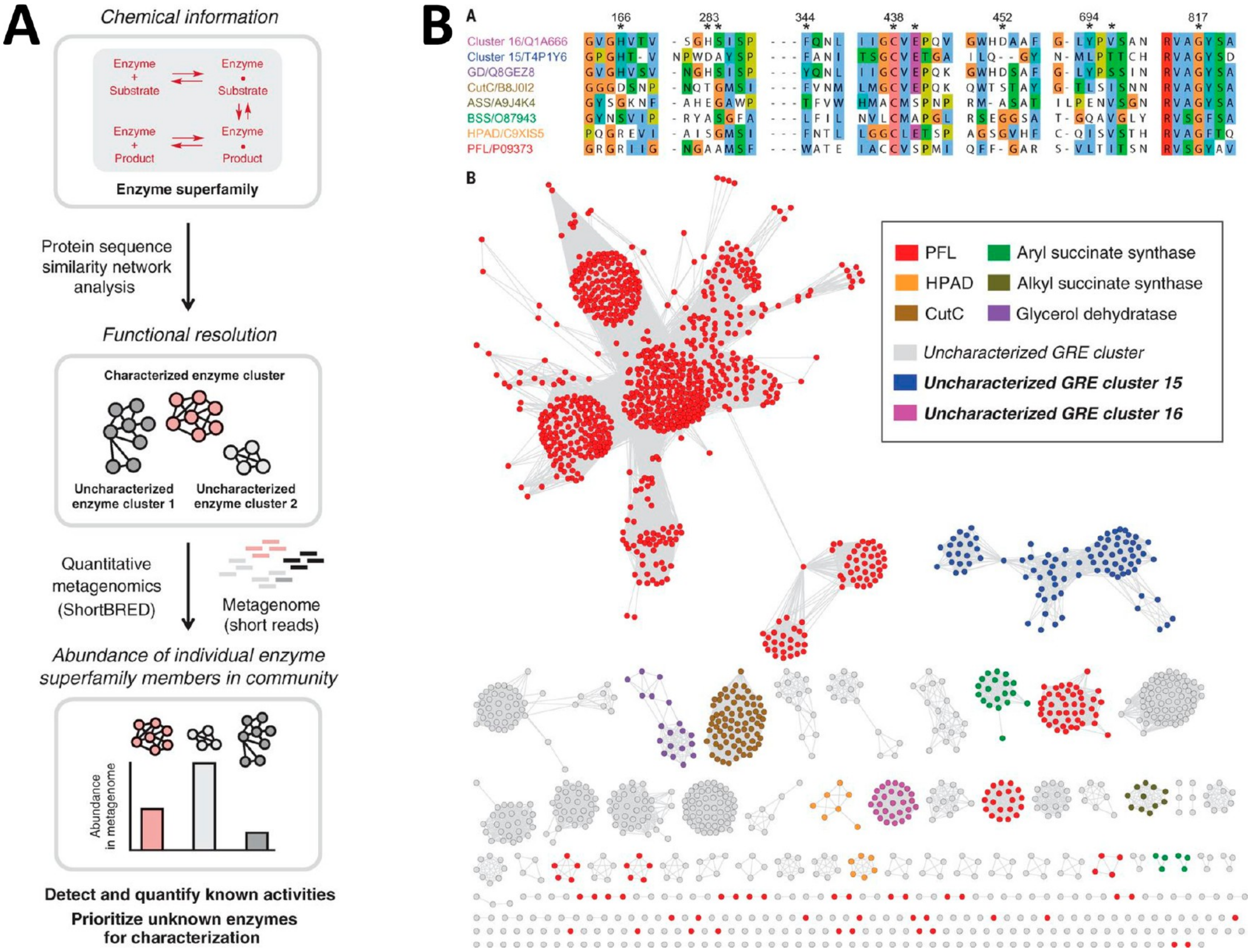

A
Chemical information
Enzyme + Substrate
Enzyme • Substrate
Enzyme + Product
Enzyme • Product
Enzyme superfamily
Protein sequence similarity network analysis
Functional resolution
Characterized enzyme cluster
Uncharacterized enzyme cluster 1
Uncharacterized enzyme cluster 2
Quantitative metagenomics (ShortBRED)
Metagenome (short reads)
Abundance of individual enzyme superfamily members in community
Abundance in metagenome
Detect and quantify known activities
Prioritize unknown enzymes for characterization
B
A
166 283 344 438 452 694 817
Cluster 16/Q1A666 GVGHVTV -SGHSISP ---FQNL IIGCVEPQV GWHDAAF G-LYPVSAN RVAGYSA
Cluster 15/T4P1Y6 GPGHT-V NPWDAYSP ---FANI TSGCVETGA ILQ--GY N-MLPTTCH RVAGYSD
GD/Q8GEZ8 GVGHVSV -NGHSISP ---YQNL IIGCVEPQK GWHDSAF G-LYPSSIN RVAGYSA
CutC/B8J0I2 GGGDSNP -NQTGMSI ---FVNM LMGCVEPQK QWTSTAY G-TLSISNN RVAGYSA
ASS/A9J4K4 GYSGKNF -AHEGAWP ---TFVW HMACMSPNP RM-ASAT ILPENVSGN RVAGYSA
BSS/O87943 GYNSVIP -RYASGFA ---LFIL NVLCMAPGL RSEGGSA T-GQAVGLY RVSGFSA
HPAD/C9XIS5 PQGREVI -AISGMSI ---FNTL LGGCLETSP AGSGVHF C-QISVSTH RVAGFTQ
PFL/P09373 GRGRIIG -NGAAMSF ---WATE IACCVSPMI QFF-GAR S-VLTITSN RVSGYAV
B
🟥 PFL
🟩 Aryl succinate synthase
HPAD
Alkyl succinate synthase
CutC
Glycerol dehydratase
Uncharacterized GRE cluster
Uncharacterized GRE cluster 15
Uncharacterized GRE cluster 16

Figure 9. (A) Strategy for chemically guided functional profiling. (B) SSN for the glycyl radical enzyme superfamily showing clusters with previously assigned functions as well as clusters (15 and 16) for which chemically guided functional profiling was used to leverage experimental functional assignment. Figures reproduced with permission from ref 72.

target selection strategy is a major challenge for functional assignment. The SSNs for functionally diverse enzyme families often have many uncharacterized clusters—the problem is deciding which are worth experimental characterization. One approach is to select those that are most biologically relevant, but how is that achieved in the absence of knowledge of their functions?

Balskus and Huttenhower recently described a strategy for choosing biologically relevant targets termed "chemically guided functional profiling".[72] This strategy involves (1) construction of the SSN for a targeted protein family segregated into isofunctional families and (2) mapping the abundance of metagenome reads to the clusters in the SSN, with uncharacterized clusters having the largest number of metagenome markers the highest priority for functional characterization (Figure 9A). ShortBRED[100] provides a fast and accurate method to profile metagenome samples and uses sequence fragments from the clusters in the SSN ("markers') to identify homologous sequences in the metagenome reads; their abundance is then mapped to the SSN clusters to accomplish target selection.

The utility of chemically guided functional profiling was demonstrated using the glycyl radical enzyme (GRE) superfamily; the reactions are initiated by abstraction of a hydrogen atom from the substrate by a glycine-centered backbone radical (generated by an activase from the *S*-adenosyl methionine superfamily). The metagenome samples used for target selection were from the human gut microbiome, so uncharacterized members of the GRE superfamily are likely involved in reactions that allow the microbiome to utilize small molecules in the gut. Balskus previously had identified choline trimethylamine-lyase (CutC) in human gut microbiome species; CutC catalyzes the cleavage of choline to acetaldehyde and trimethylamine, the latter involved in the production of methane as well as implicated in human diseases via its *N*-oxide.[101,102]

The SSN for the GRE family is shown in Figure 9B. The functionally assigned clusters are colored, as are two clusters (15 and 16) that were identified as abundant in the human gut microbiome. Both of the latter clusters were hypothesized to be dehydratases based on conserved active site residues associated with known dehydratase reactions. Cluster 15 was characterized as a 4-hydroxyproline dehydratase; again, genome context was used to predict the substrate because of its proximity to Δ^1-pyrroline-5-carboxylate (P5C) reductase that reduces P5C that would be derived from dehydration of 4-hydroxyproline to proline. Cluster 16 was characterized as a novel (*S*)-1,2-

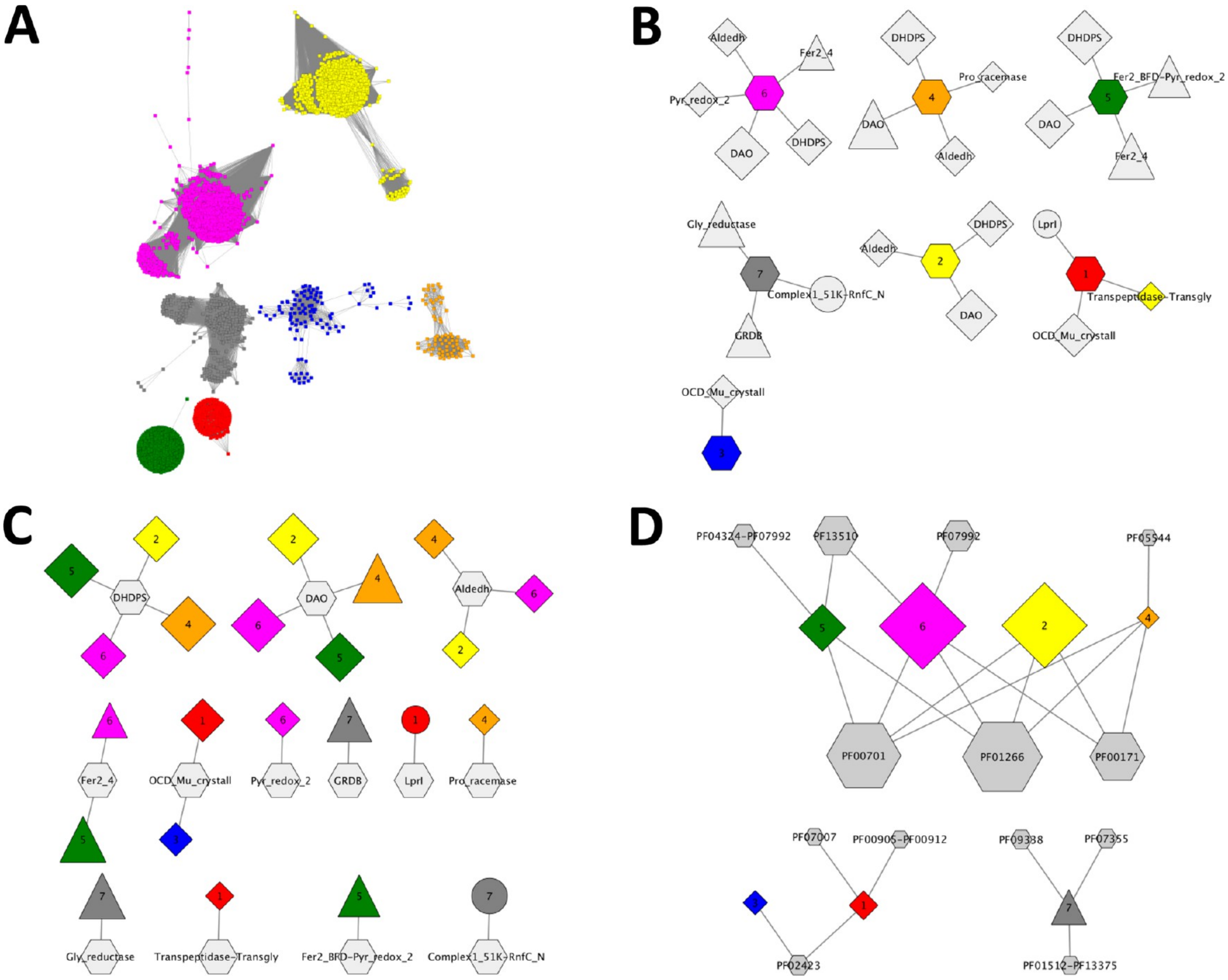


Figure 10. (A) Colored SSN generated by EFI-GNT for selected clusters in the proline racemase family (PF05544). (B) GNN with SSN cluster hub-nodes and Pfam family spoke-nodes. (C) GNN with Pfam family hub-nodes and SSN cluster spoke-nodes. (D) Refined GNN showing identification of three different functions as deduced by connections (or lack thereof) between SSN cluster and Pfam family nodes.

propanediol dehydratase (a previously characterized analogue is an adenosylcobalamin-dependent enzyme); the identity of the substrate was suggested from genome analysis because the enzyme is found in *Roseburia inulinivorans* that catabolizes L-fucose but lacks the adenosylcobalamin-dependent dehydratase.

A "user friendly" web tool is not yet available to allow the community to use "chemically guided functional profiling" with their favorite families. But, the development of a web tool is a high priority goal given its ability to identify important targets for functional characterization.

AGeNNT and Refined GNNs: Building on EFI-GNT. EFI-GNT provides GNNs in two formats that summarize (1) the Pfam families identified by each SSN cluster (edges between SSN cluster hub-nodes and Pfam family spoke-nodes), providing information about the reactions in metabolic pathways, and (2) the SSN clusters that identify each Pfam family (edges between Pfam family hub-nodes and SSN cluster spoke-nodes), providing information about whether multiple clusters may contain orthologues.

Merkl and co-workers recently described AGeNNT (Automatically Generates refined Neighborhood NeTworks), a Java application that uses the GNNs provided by EFI-GNT to generate a third format ("refined GNN") in which all of the SSN cluster and Pfam family nodes are connected by edges.[71] Clusters that contain orthologues, identified when they share the same genome neighbors, can be distinguished from clusters that have different genome contexts. An SSN is submitted to the EFI-GNT web tool. AGeNNT then generates the refined GNN. Several options are provided, including (1) eliminating overrepresented phylogenetically related subspecies from the input SSN to reduce redundancy in the GNN and (2) using a user-defined "whitelist" of Pfam families to include in the refined GNN. For example, only Pfam families for enzymes can be included in the refined GNN so Pfam cluster connections between SSN clusters that involve transporters and transcriptional regulators are eliminated (in contrast to pathway enzymes, transporters and transcriptional regulators are not conserved).

Continuing again with the proline racemase family (PF05544) to provide an example, several major clusters from the SSN were selected for generation of GNNs using EFI-GNT and the refined GNN using AGeNNT (Figure 10). The colored SSN is shown in Figure 10A, the SSN cluster hub-node GNN format is shown in Figure 10B, the Pfam family hub-node GNN format is shown in Figure 10C, and the refined GNN is shown in Figure 10D (Pfam families for transport systems and transcriptional regulators are deleted in the GNNs; because these families are not conserved in pathways (*vide supra*), their inclusion in the refined GNN can complicate the analysis). Comparison of the refined GNN with the GNNs establishes the utility of the refined GNN in identifying orthologous SSN clusters: clusters 2, 4, 5, and 6 are orthologous 4-hydroxyproline epimerases; clusters 1 and 3 are orthologous *trans*-3-hydroxylproline dehydratases; and cluster 7 is proline racemase (using functional assignments based on experimental verification[97]). Building on EFI-EST and EFI-GNT, AGeNNT links SSN clusters that share pathway context, potentially identifying interrelations of subfamilies within a protein family.

Future Directions. EFI-EST and EFI-GNT provide experimentalists with otherwise inaccessible but essential perspectives on sequence−function space in protein families and genome context that facilitate the assignment of functions to uncharacterized enzymes. Other web tools are available for smaller scale analysis of protein families, but genomic enzymology "requires" large-scale analyses to provide the maximum amount of context.

Other large-scale web tools can be imagined. For example, the proteome of an organism (or of a community) determines its metabolic capabilities; therefore, an easy-to-construct overview of the metabolic potential would be useful and could be provided by a "proteome network" (PN) tool. A PN would include a node for each protein encoded by a genome (or community) and collected into Pfam family clusters (Pfam family hub-node and protein spoke nodes). The PN would identify the catalytic capabilities via the identities of the Pfam families and, also, the locations of the proteins (spoke nodes) in the SSNs for their families. For a community PN, identification of species-specific Pfam families could provide the potential to identify syntrophic metabolic pathways, e.g., different organisms contribute different metabolic capabilities to synthesize a natural product or degrade an energy source. In analogy with chemically guided functional profiling, mapping transcriptome abundance to the PN would provide a visually powerful approach for identifying enzymes in novel pathways.

Also, the Pfam families that contribute enzymes to a pathway often are conserved in phylogenetically diverse organisms; however, we have observed that one or more reactions in a metabolic pathway can be catalyzed by analogues (non-orthologous gene replacements) in different taxonomic ranks, e.g., phyla, class, order, or family. The ability to discover analogues may be enhanced by clustering members of a protein family by taxonomic rank instead of pairwise sequence identity (SSNs). Because the node attributes that are provided by EFI-EST for sequences include taxonomic ranking, a taxonomic rank network ("TRN") would be easy to construct. Subsequent generation of sequence similarity-based SSNs for individual clusters in the TRN would be accomplished with Option D of EFI-EST, thereby providing the ability to further segregate and analyze the clusters by sequence homology.

Finally, although the generation of an SSN is straightforward, Release 31.0 of the Pfam database (Release 31.0) defines 16 712 families. Immediate access to a library of precomputed SSNs for all Pfam families would provide the biological and biomedical communities, including users of web tools that identify BGCs (*vide supra*), with the ability to quickly place their favorite enzymes in the context sequence−function relationships for their protein families. This library of SSNs should be regularly updated to provide current information (perhaps in parallel with releases of the InterPro database), but its construction requires considerable computational resources. We have demonstrated that the calculation of this database is feasible, although we have not yet been able to initiate the production phase of this effort.

I encourage the readers to (1) try the EFI-EST and EFI-GNT web tools, (2) imagine new applications for SSNs and GNNs, and (3) identify additional large-scale data visualization and analysis challenges that would be amenable to solution by community-accessible web tools. Like the natural products community, the enzymology community needs to recognize the essential role of web tools that allow the protein and genome sequence databases to be leveraged for the solution of biological problems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: j-gerlt@uiuc.edu.

ORCID

John A. Gerlt: 0000-0002-5625-1218

Funding

This work was supported by U54GM093442 and P01GM118303 from the National Institutes of Health.

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

I thank Mr. Daniel Davidson and Mr. Nils Oberg for writing the software that supports EFI-EST and EFI-GNT. I also thank Drs. Tyler Stack and Rémi Zallot for their critical comments on the manuscript.

REFERENCES

(1) Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem. 70*, 209−246.

(2) Furnham, N., Dawson, N. L., Rahman, S. A., Thornton, J. M., and Orengo, C. A. (2016) Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies. *J. Mol. Biol. 428*, 253−267.

(3) Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Goker, M., Coates, R. C., Hadjithomas, M., Pavlopoulos, G. A., Paez-Espino, D., Yoshikuni, Y., Visel, A., Whitman, W. B., Garrity, G. M., Eisen, J. A., Hugenholtz, P., Pati, A., Ivanova, N. N., Woyke, T., Klenk, H. P., and Kyrpides, N. C. (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol. 35*, 676−683.

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

(5) Khosla, C. (2015) Quo vadis, enzymology? *Nat. Chem. Biol. 11*, 438−441.

(6) Gerlt, J. A., Allen, K. N., Almo, S. C., Armstrong, R. N., Babbitt, P. C., Cronan, J. E., Dunaway-Mariano, D., Imker, H. J., Jacobson, M. P., Minor, W., Poulter, C. D., Raushel, F. M., Sali, A., Shoichet, B. K., and Sweedler, J. V. (2011) The Enzyme Function Initiative. *Biochemistry 50*, 9950−9962.

(7) Ikeda, H., Nonomiya, T., Usami, M., Ohta, T., and Omura, S. (1999) Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in Streptomyces avermitilis. *Proc. Natl. Acad. Sci. U. S. A. 96*, 9509−9514.

(8) Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth,

S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J., and Hopwood, D. A. (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature 417*, 141−147.

(9) Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., and Omura, S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat. Biotechnol. 21*, 526−531.

(10) Yu, X., Doroghazi, J. R., Janga, S. C., Zhang, J. K., Circello, B., Griffin, B. M., Labeda, D. P., and Metcalf, W. W. (2013) Diversity and abundance of phosphonate biosynthetic genes in nature. *Proc. Natl. Acad. Sci. U. S. A. 110*, 20759−20764.

(11) Ju, K. S., Gao, J., Doroghazi, J. R., Wang, K. K., Thibodeaux, C. J., Li, S., Metzger, E., Fudala, J., Su, J., Zhang, J. K., Lee, J., Cioni, J. P., Evans, B. S., Hirota, R., Labeda, D. P., van der Donk, W. A., and Metcalf, W. W. (2015) Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc. Natl. Acad. Sci. U. S. A. 112*, 12175−12180.

(12) Medema, M. H., and Fischbach, M. A. (2015) Computational approaches to natural product discovery. *Nat. Chem. Biol. 11*, 639−648.

(13) Tietz, J. I., and Mitchell, D. A. (2016) Using Genomics for Natural Product Structure Elucidation. *Curr. Top. Med. Chem. 16*, 1645−1694.

(14) Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., Suarez Duran, H. G., de Los Santos, E. L. C., Kim, H. U., Nave, M., Dickschat, J. S., Mitchell, D. A., Shelest, E., Breitling, R., Takano, E., Lee, S. Y., Weber, T., and Medema, M. H. (2017) antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res. 45*, W36−W41.

(15) Skinnider, M. A., Merwin, N. J., Johnston, C. W., and Magarvey, N. A. (2017) PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res. 45*, W49−W54.

(16) Tietz, J. I., Schwalen, C. J., Patel, P. S., Maxson, T., Blair, P. M., Tai, H. C., Zakai, U. I., and Mitchell, D. A. (2017) A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol. 13*, 470−478.

(17) Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, 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., and Babbitt, P. C. (2014) The Structure-Function Linkage Database. *Nucleic Acids Res. 42*, D521−530.

(18) Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry 35*, 16489−16501.

(19) Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. (2012) Divergent evolution in enolase superfamily: strategies for assigning functions. *J. Biol. Chem. 287*, 29−34.

(20) Schmidt, D. M., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J., and Gerlt, J. A. (2003) Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry 42*, 8387−8393.

(21) Vick, J. E., Schmidt, D. M., and Gerlt, J. A. (2005) Evolutionary potential of (beta/alpha)8-barrels: in vitro enhancement of a "new" reaction in the enolase superfamily. *Biochemistry 44*, 11722−11729.

(22) Engelhardt, B. E., Jordan, M. I., Repo, S. T., and Brenner, S. E. (2009) Phylogenetic molecular function annotation. *J. Phys.: Conf. Ser. 180*, 012024.

(23) Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science 324*, 1561−1564.

(24) Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) FastTree 2– approximately maximum-likelihood trees for large alignments. *PLoS One 5*, e9490.

(25) Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One 4*, e4345.

(26) Kohl, M., Wiese, S., and Warscheid, B. (2011) Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol. 696*, 291−303.

(27) Brown, S. D., Gerlt, J. A., Seffernick, J. L., and Babbitt, P. C. (2006) A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol. 7*, R8.

(28) Barber, A. E., 2nd, and Babbitt, P. C. (2012) Pythoscape: a framework for generation of large protein similarity networks. *Bioinformatics 28*, 2845−2846.

(29) Li, W., Kinch, L. N., and Grishin, N. V. (2013) Pclust: protein network visualization highlighting experimental data. *Bioinformatics 29*, 2647−2648.

(30) Frickey, T., and Lupas, A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics 20*, 3702−3704.

(31) Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., and Whalen, K. L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta, Proteins Proteomics 1854*, 1019−1037.

(32) Colin, P. Y., Kintses, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., Hyvonen, M., Morgavi, D. P., Janssen, D. B., and Hollfelder, F. (2015) Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun. 6*, 10008.

(33) Cox, C. L., Doroghazi, J. R., and Mitchell, D. A. (2015) The genomic landscape of ribosomal peptides containing thiazole and oxazole heterocycles. *BMC Genomics 16*, 778.

(34) Huang, H., Carter, M. S., Vetting, M. W., Al-Obaidi, N., Patskovsky, Y., Almo, S. C., and Gerlt, J. A. (2015) A General Strategy for the Discovery of Metabolic Pathways: d-Threitol, l-Threitol, and Erythritol Utilization in Mycobacterium smegmatis. *J. Am. Chem. Soc. 137*, 14570−14573.

(35) Petronikolou, N., and Nair, S. K. (2015) Biochemical Studies of Mycobacterial Fatty Acid Methyltransferase: A Catalyst for the Enzymatic Production of Biodiesel. *Chem. Biol. 22*, 1480−1490.

(36) Rao, G., O'Dowd, B., Li, J., Wang, K., and Oldfield, E. (2015) IspH-RPS1 and IspH-UbiA: "Rosetta Stone" Proteins. *Chem. Sci. 6*, 6813−6822.

(37) Roche, D. B., Brackenridge, D. A., and McGuffin, L. J. (2015) Proteins and Their Interacting Partners: An Introduction to Protein-Ligand Binding Site Prediction Methods. *Int. J. Mol. Sci. 16*, 29829−29842.

(38) Wichelecki, D. J., Vetting, M. W., Chou, L., Al-Obaidi, N., Bouvier, J. T., Almo, S. C., and Gerlt, J. A. (2015) ATP-binding Cassette (ABC) Transport System Solute-binding Protein-guided Identification of Novel d-Altritol and Galactitol Catabolic Pathways in Agrobacterium tumefaciens C58. *J. Biol. Chem. 290*, 28963−28976.

(39) Ahmed, F. H., Mohamed, A. E., Carr, P. D., Lee, B. M., Condic-Jurkic, K., O'Mara, M. L., and Jackson, C. J. (2016) Rv2074 is a novel F420 H2-dependent biliverdin reductase in Mycobacterium tuberculosis. *Protein Sci. 25*, 1692−1709.

(40) Atkinson, J. T., Campbell, I., Bennett, G. N., and Silberg, J. J. (2016) Cellular Assays for Ferredoxins: A Strategy for Understanding Electron Flow through Protein Carriers That Link Metabolic Pathways. *Biochemistry 55*, 7047−7064.

(41) Baier, F., Copp, J. N., and Tokuriki, N. (2016) Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence-Function Relationships. *Biochemistry 55*, 6375−6388.

(42) Bhandari, D. M., Fedoseyenko, D., and Begley, T. P. (2016) Tryptophan Lyase (NosL): A Cornucopia of 5′-Deoxyadenosyl Radical Mediated Transformations. *J. Am. Chem. Soc. 138*, 16184−16187.

(43) Chekan, J. R., Koos, J. D., Zong, C., Maksimov, M. O., Link, A. J., and Nair, S. K. (2016) Structure of the Lasso Peptide Isopeptidase Identifies a Topology for Processing Threaded Substrates. *J. Am. Chem. Soc. 138*, 16452−16458.

(44) Davey, L., Halperin, S. A., and Lee, S. F. (2016) Thiol-Disulfide Exchange in Gram-Positive Firmicutes. *Trends Microbiol. 24*, 902−915.

(45) Desai, J., Liu, Y. L., Wei, H., Liu, W., Ko, T. P., Guo, R. T., and Oldfield, E. (2016) Structure, Function, and Inhibition of Staphylococcus aureus Heptaprenyl Diphosphate Synthase. *ChemMedChem 11*, 1915−1923.

(46) Ding, W., Li, Q., Jia, Y., Ji, X., Qianzhu, H., and Zhang, Q. (2016) Emerging Diversity of the Cobalamin-Dependent Methyltransferases Involving Radical-Based Mechanisms. *ChemBioChem 17*, 1191−1197.

(47) Gerlt, J. A. (2016) Tools and strategies for discovering novel enzymes and metabolic pathways. *Perspectives in Science 9*, 24−32.

(48) Ghodge, S. V., Biernat, K. A., Bassett, S. J., Redinbo, M. R., and Bowers, A. A. (2016) Post-translational Claisen Condensation and Decarboxylation en Route to the Bicyclic Core of Pantocin A. *J. Am. Chem. Soc. 138*, 5487−5490.

(49) Hao, Y., Pierce, E., Roe, D., Morita, M., McIntosh, J. A., Agarwal, V., Cheatham, T. E., 3rd, Schmidt, E. W., and Nair, S. K. (2016) Molecular basis for the broad substrate selectivity of a peptide prenyltransferase. *Proc. Natl. Acad. Sci. U. S. A. 113*, 14037−14042.

(50) Ji, X., Li, Y., Xie, L., Lu, H., Ding, W., and Zhang, Q. (2016) Expanding Radical SAM Chemistry by Using Radical Addition Reactions and SAM Analogues. *Angew. Chem., Int. Ed. 55*, 11845−11848.

(51) Ji, X., Liu, W. Q., Yuan, S., Yin, Y., Ding, W., and Zhang, Q. (2016) Mechanistic study of the radical SAM-dependent amine dehydrogenation reactions. *Chem. Commun. 52*, 10555−10558.

(52) Kumar, G., Johnson, J. L., and Frantom, P. A. (2016) Improving Functional Annotation in the DRE-TIM Metallolyase Superfamily through Identification of Active Site Fingerprints. *Biochemistry 55*, 1863−1872.

(53) Li, D., Moorman, R., Vanhercke, T., Petrie, J., Singh, S., and Jackson, C. J. (2016) Classification and substrate head-group specificity of membrane fatty acid desaturases. *Comput. Struct. Biotechnol. J. 14*, 341−349.

(54) Molloy, E. M., Tietz, J. I., Blair, P. M., and Mitchell, D. A. (2016) Biological characterization of the hygrobafilomycin antibiotic JBIR-100 and bioinformatic insights into the hygrolide family of natural products. *Bioorg. Med. Chem. 24*, 6276−6290.

(55) Plach, M. G., Reisinger, B., Sterner, R., and Merkl, R. (2016) Long-Term Persistence of Bi-functionality Contributes to the Robustness of Microbial Life through Exaptation. *PLoS Genet. 12*, e1005836.

(56) Rao, G., and Oldfield, E. (2016) Structure and Function of Four Classes of the 4Fe-4S Protein, IspH. *Biochemistry 55*, 4119−4129.

(57) Thotsaporn, K., Tinikul, R., Maenpuen, S., Phonbuppha, J., Watthaisong, P., Chenprakhon, P., and Chaiyen, P. (2016) Enzymes in the p-hydroxphenylacetate degradation pathway of Acinetobacter baumannii. *J. Mol. Catal. B: Enzym. 134*, 353−366.

(58) Zallot, R., Harrison, K. J., Kolaczkowski, B., and de Crecy-Lagard, V. (2016) Functional Annotations of Paralogs: A Blessing and a Curse. *Life 6*, 39.

(59) Zhang, X., Carter, M. S., Vetting, M. W., San Francisco, B., Zhao, S., Al-Obaidi, N. F., Solbiati, J. O., Thiaville, J. J., de Crecy-Lagard, V., Jacobson, M. P., Almo, S. C., and Gerlt, J. A. (2016) Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars. *Proc. Natl. Acad. Sci. U. S. A. 113*, E4161−4169.

(60) Ahmed, M. N., Reyna-Gonzalez, E., Schmid, B., Wiebach, V., Sussmuth, R. D., Dittmann, E., and Fewer, D. P. (2017) Phylogenomic Analysis of the Microviridin Biosynthetic Pathway Coupled with Targeted Chemo-Enzymatic Synthesis Yields Potent Protease Inhibitors. *ACS Chem. Biol. 12*, 1538.

(61) Bearne, S. L. (2017) The interdigitating loop of the enolase superfamily as a specificity binding determinant or 'flying buttress'. *Biochim. Biophys. Acta, Proteins Proteomics 1865*, 619−630.

(62) Benjdia, A., Guillot, A., Ruffie, P., Leprince, J., and Berteau, O. (2017) Post-translational modification of ribosomally synthesized peptides by a radical SAM epimerase in Bacillus subtilis. *Nat. Chem. 9*, 698−707.

(63) Erb, T. J., Jones, P. R., and Bar-Even, A. (2017) Synthetic metabolism: metabolic engineering meets enzyme design. *Curr. Opin. Chem. Biol. 37*, 56−62.

(64) Estrada, P., Manandhar, M., Dong, S. H., Deveryshetty, J., Agarwal, V., Cronan, J. E., and Nair, S. K. (2017) The pimeloyl-CoA synthetase BioW defines a new fold for adenylate-forming enzymes. *Nat. Chem. Biol. 13*, 668−674.

(65) Giessen, T. W., and Silver, P. A. (2017) Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nat. Microbiol 2*, 17029.

(66) Glasner, M. E. (2017) Finding enzymes in the gut metagenome. *Science 355*, 577−578.

(67) Hetrick, K. J., and van der Donk, W. A. (2017) Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era. *Curr. Opin. Chem. Biol. 38*, 36−44.

(68) Holliday, G. L., Brown, S. D., Akiva, E., Mischel, D., Hicks, M. A., Morris, J. H., Huang, C. C., Meng, E. C., Pegg, S. C., Ferrin, T. E., and Babbitt, P. C. (2017) Biocuration in the structure-function linkage database: the anatomy of a superfamily. *Database*, DOI: 10.1093/database/bax045.

(69) Jia, B., Jia, X., Hyun Kim, K., Ji Pu, Z., Kang, M. S., and Ok Jeon, C. (2017) Evolutionary, computational, and biochemical studies of the salicylaldehyde dehydrogenases in the naphthalene degradation pathway. *Sci. Rep. 7*, 43489.

(70) Jia, B., Jia, X., Kim, K. H., and Jeon, C. O. (2017) Integrative view of 2-oxoglutarate/Fe(II)-dependent oxygenase diversity and functions in bacteria. *Biochim. Biophys. Acta, Gen. Subj. 1861*, 323−334.

(71) Kandlinger, F., Plach, M. G., and Merkl, R. (2017) AGeNNT: annotation of enzyme families by means of refined neighborhood networks. *BMC Bioinf. 18*, 274.

(72) Levin, B. J., Huang, Y. Y., Peck, S. C., Wei, Y., Martinez-Del Campo, A., Marks, J. A., Franzosa, E. A., Huttenhower, C., and Balskus, E. P. (2017) A prominent glycyl radical enzyme in human gut microbiomes metabolizes trans-4-hydroxy-l-proline. *Science 355*, eaai8386.

(73) Ney, B., Ahmed, F. H., Carere, C. R., Biswas, A., Warden, A. C., Morales, S. E., Pandey, G., Watt, S. J., Oakeshott, J. G., Taylor, M. C., Stott, M. B., Jackson, C. J., and Greening, C. (2017) The methanogenic redox cofactor F420 is widely synthesized by aerobic soil bacteria. *ISME J. 11*, 125−137.

(74) Ortega, M. A., Cogan, D. P., Mukherjee, S., Garg, N., Li, B., Thibodeaux, G. N., Maffioli, S. I., Donadio, S., Sosio, M., Escano, J., Smith, L., Nair, S. K., and van der Donk, W. A. (2017) Two Flavoenzymes Catalyze the Post-Translational Generation of 5-Chlorotryptophan and 2-Aminovinyl-Cysteine during NAI-107 Biosynthesis. *ACS Chem. Biol. 12*, 548−557.

(75) Pimviriyakul, P., Thotsaporn, K., Sucharitakul, J., and Chaiyen, P. (2017) Kinetic Mechanism of the Dechlorinating Flavin-dependent Monooxygenase HadA. *J. Biol. Chem. 292*, 4818−4832.

(76) Repka, L. M., Chekan, J. R., Nair, S. K., and van der Donk, W. A. (2017) Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem. Rev. 117*, 5457−5520.

(77) Schwalen, C. J., Feng, X., Liu, W., O-Dowd, B., Ko, T. P., Shin, C. J., Guo, R. T., Mitchell, D. A., and Oldfield, E. (2017) Head-to-Head Prenyl Synthases in Pathogenic Bacteria. *ChemBioChem 18*, 985−991.

(78) Zallot, R., Yuan, Y., and de Crecy-Lagard, V. (2017) The *Escherichia coli* COG1738 Member YhhQ Is Involved in 7-Cyanodeazaguanine (preQ(0)) Transport. *Biomolecules 7*, 12.

(79) Xiang, D. F., Kolb, P., Fedorov, A. A., Xu, C., Fedorov, E. V., Narindoshivili, T., Williams, H. J., Shoichet, B. K., Almo, S. C., and Raushel, F. M. (2012) Structure-based function discovery of an enzyme for the hydrolysis of phosphorylated sugar lactones. *Biochemistry 51*, 1762−1773.

(80) Fan, H., Hitchcock, D. S., Seidel, R. D., 2nd, Hillerich, B., Lin, H., Almo, S. C., Sali, A., Shoichet, B. K., and Raushel, F. M. (2013) Assignment of pterin deaminase activity to an enzyme of unknown function guided by homology modeling and docking. *J. Am. Chem. Soc. 135*, 795−803.

(81) Goble, A. M., Toro, R., Li, X., Ornelas, A., Fan, H., Eswaramoorthy, S., Patskovsky, Y., Hillerich, B., Seidel, R., Sali, A., Shoichet, B. K., Almo, S. C., Swaminathan, S., Tanner, M. E., and Raushel, F. M. (2013) Deamination of 6-aminodeoxyfutalosine in menaquinone biosynthesis by distantly related enzymes. *Biochemistry 52*, 6525−6536.

(82) Hitchcock, D. S., Fan, H., Kim, J., Vetting, M., Hillerich, B., Seidel, R. D., Almo, S. C., Shoichet, B. K., Sali, A., and Raushel, F. M. (2013) Structure-guided discovery of new deaminase enzymes. *J. Am. Chem. Soc. 135*, 13927−13933.

(83) Ornelas, A., Korczynska, M., Ragumani, S., Kumaran, D., Narindoshvili, T., Shoichet, B. K., Swaminathan, S., and Raushel, F. M. (2013) Functional annotation and three-dimensional structure of an incorrectly annotated dihydroorotase from cog3964 in the amidohydrolase superfamily. *Biochemistry 52*, 228−238.

(84) Barelier, S., Cummings, J. A., Rauwerdink, A. M., Hitchcock, D. S., Farelli, J. D., Almo, S. C., Raushel, F. M., Allen, K. N., and Shoichet, B. K. (2014) Substrate deconstruction and the nonadditivity of enzyme recognition. *J. Am. Chem. Soc. 136*, 7374−7382.

(85) Korczynska, M., Xiang, D. F., Zhang, Z., Xu, C., Narindoshvili, T., Kamat, S. S., Williams, H. J., Chang, S. S., Kolb, P., Hillerich, B., Sauder, J. M., Burley, S. K., Almo, S. C., Swaminathan, S., Shoichet, B. K., and Raushel, F. M. (2014) Functional annotation and structural characterization of a novel lactonase hydrolyzing D-xylono-1,4-lactone-5-phosphate and L-arabino-1,4-lactone-5-phosphate. *Biochemistry 53*, 4727−4738.

(86) Lukk, T., Sakai, A., Kalyanaraman, C., Brown, S. D., Imker, H. J., Song, L., Fedorov, A. A., Fedorov, E. V., Toro, R., Hillerich, B., Seidel, R., Patskovsky, Y., Vetting, M. W., Nair, S. K., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc. Natl. Acad. Sci. U. S. A. 109*, 4122−4127.

(87) Wichelecki, D. J., Balthazor, B. M., Chau, A. C., Vetting, M. W., Fedorov, A. A., Fedorov, E. V., Lukk, T., Patskovsky, Y. V., Stead, M. B., Hillerich, B. S., Seidel, R. D., Almo, S. C., and Gerlt, J. A. (2014) Discovery of function in the enolase superfamily: D-mannonate and d-gluconate dehydratases in the D-mannonate dehydratase subgroup. *Biochemistry 53*, 2722−2731.

(88) Wichelecki, D. J., Froese, D. S., Kopec, J., Muniz, J. R., Yue, W. W., and Gerlt, J. A. (2014) Enzymatic and structural characterization of rTSgamma provides insights into the function of rTSbeta. *Biochemistry 53*, 2732−2738.

(89) Wichelecki, D. J., Graff, D. C., Al-Obaidi, N., Almo, S. C., and Gerlt, J. A. (2014) Identification of the in vivo function of the high-efficiency D-mannonate dehydratase in Caulobacter crescentus NA1000 from the enolase superfamily. *Biochemistry 53*, 4087−4089.

(90) Wichelecki, D. J., Vendiola, J. A., Jones, A. M., Al-Obaidi, N., Almo, S. C., and Gerlt, J. A. (2014) Investigating the physiological roles of low-efficiency D-mannonate and D-gluconate dehydratases in the enolase superfamily: pathways for the catabolism of L-gulonate and L-idonate. *Biochemistry 53*, 5692−5699.

(91) Ghasempur, S., Eswaramoorthy, S., Hillerich, B. S., Seidel, R. D., Swaminathan, S., Almo, S. C., and Gerlt, J. A. (2014) Discovery of a novel L-lyxonate degradation pathway in Pseudomonas aeruginosa PAO1. *Biochemistry 53*, 3357−3366.

(92) Groninger-Poe, F. P., Bouvier, J. T., Vetting, M. W., Kalyanaraman, C., Kumar, R., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2014) Evolution of enzymatic activities in the enolase superfamily: galactarate dehydratase III from Agrobacterium tumefaciens C58. *Biochemistry 53*, 4192−4203.

(93) Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., Jagessar, K., Kim, J., Patskovsky, Y., Seidel, R. D., Stead, M., Toro, R., Vetting, M. W., Almo, S. C., Armstrong, R. N., and Babbitt, P. C. (2014) Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol. 12*, e1001843.

(94) Huang, H., Pandya, C., Liu, C., Al-Obaidi, N. F., Wang, M., Zheng, L., Toews Keating, S., Aono, M., Love, J. D., Evans, B., Seidel, R. D., Hillerich, B. S., Garforth, S. J., Almo, S. C., Mariano, P. S., Dunaway-Mariano, D., Allen, K. N., and Farelli, J. D. (2015) Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc. Natl. Acad. Sci. U. S. A. 112*, E1974−1983.

(95) Tian, B. X., Wallrapp, F. H., Holiday, G. L., Chow, J. Y., Babbitt, P. C., Poulter, C. D., and Jacobson, M. P. (2014) Predicting the functions and specificity of triterpenoid synthases: a mechanism-based multi-intermediate docking approach. *PLoS Comput. Biol. 10*, e1003874.

(96) Wallrapp, F. H., Pan, J. J., Ramamoorthy, G., Almonacid, D. E., Hillerich, B. S., Seidel, R., Patskovsky, Y., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Poulter, C. D. (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proc. Natl. Acad. Sci. U. S. A. 110*, E1196−1202.

(97) Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., San Francisco, B., Solbiati, J., Steves, A., Brown, S., Akiva, E., Barber, A., Seidel, R. D., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2014) Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife 3*, No. e03275, DOI: 10.7554/eLife.03275.

(98) Vetting, M. W., Al-Obaidi, N., Zhao, S., San Francisco, B., Kim, J., Wichelecki, D. J., Bouvier, J. T., Solbiati, J. O., Vu, H., Zhang, X., Rodionov, D. A., Love, J. D., Hillerich, B. S., Seidel, R. D., Quinn, R. J., Osterman, A. L., Cronan, J. E., Jacobson, M. P., Gerlt, J. A., and Almo, S. C. (2015) Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry 54*, 909−931.

(99) Ndeh, D., Rogowski, A., Cartmell, A., Luis, A. S., Basle, A., Gray, J., Venditto, I., Briggs, J., Zhang, X., Labourel, A., Terrapon, N., Buffetto, F., Nepogodiev, S., Xiao, Y., Field, R. A., Zhu, Y., O'Neill, M. A., Urbanowicz, B. R., York, W. S., Davies, G. J., Abbott, D. W., Ralet, M. C., Martens, E. C., Henrissat, B., and Gilbert, H. J. (2017) Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature 544*, 65−70.

(100) Kaminski, J., Gibson, M. K., Franzosa, E. A., Segata, N., Dantas, G., and Huttenhower, C. (2015) High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput. Biol. 11*, e1004557.

(101) Craciun, S., and Balskus, E. P. (2012) Microbial conversion of choline to trimethylamine requires a glycyl radical enzyme. *Proc. Natl. Acad. Sci. U. S. A. 109*, 21307−21312.

(102) Craciun, S., Marks, J. A., and Balskus, E. P. (2014) Characterization of choline trimethylamine-lyase expands the chemistry of glycyl radical enzymes. *ACS Chem. Biol. 9*, 1408−1413.