Perspectives in Science (2016) 9, 24—32

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc

Tools and strategies for discovering novel enzymes and metabolic pathways☆

John A. Gerlt*

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

Received 13 July 2016; accepted 26 July 2016
Available online 13 September 2016

KEYWORDS
Enzyme Function Initiative;
Sequence similarity networks;
Genome neighbourhood networks;
Genomic enzymology;
Transport system solute binding proteins;
Ethanolamine catabolism

Summary The number of entries in the sequence databases continues to increase exponentially — the UniProt database is increasing with a doubling time of ~4 years (2% increase/month). Approximately 50% of the entries have uncertain, unknown, or incorrect function annotations because these are made by automated methods based on sequence homology. If the potential in complete genome sequences is to be realized, strategies and tools must be developed to facilitate experimental assignment of functions to uncharacterized proteins discovered in genome projects. The Enzyme Function Initiative (EFI; previously supported by U54GM093342 from the National Institutes of Health, now supported by P01GM118303) developed web tools for visualizing and analyzing (1) sequence and function space in protein families (EFI-EST) and (2) genome neighbourhoods in microbial and fungal genomes (EFI-GNT) to assist the design of experimental strategies for discovering the *in vitro* activities and *in vivo* metabolic functions of uncharacterized enzymes. The EFI developed an experimental platform for large-scale production of the solute binding proteins (SBPs) for ABC, TRAP, and TCT transport systems and their screening with a physical ligand library to identify the identities of the ligands for these transport systems. Because the genes that encode transport systems are often co-located with the genes that encode the catabolic pathways for the transported solutes, the identity of the SBP ligand together with the EFI-EST and EFI-GNT web tools can be used to discover new enzyme functions and new metabolic pathways. This approach is demonstrated with the characterization of a novel pathway for ethanolamine catabolism.

* Corresponding author at: Institute for Genomic Biology and Departments of Biochemistry and Chemistry, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States.
E-mail address: j-gerlt@illinois.edu

http://dx.doi.org/10.1016/j.pisc.2016.07.001

Introduction

More than 20 years have passed since the first microbial genome sequencing project, for *Haemophilus influenza* Rd, was completed, providing the complete set of sequences for the 1707 proteins encoded by its genome (Fleischmann et al., 1995). Now, the number of protein sequences in the UniProt database (http://www.uniprot.org/) exceeds 60M and is increasing at the rate of 2%/month (Fig. 1; 4 years doubling time), the majority of the sequences are obtained from microbial genome projects. This explosion in the number of protein sequences provides the potential for discovering novel enzymes in novel metabolic pathways, a boon to enzymologists, chemical biologists, microbiologists, and systems biologists. However, perhaps 50% of the proteins in the databases have incorrect, uncertain, or unknown functions. Therefore, an important challenge for enzymology is to devise tools for mining the databases for novel enzymes and experimental strategies for determining their *in vitro* activities and *in vivo* metabolic/physiological functions.

Indeed, Dr. Chaitan Khosla recently identified this challenge as one of the major challenges, and opportunities, for contemporary enzymology (Khosla, 2015):

> ''**Enzyme function annotation.** It has long been appreciated that assigning function to enzymes based on sequence alone is difficult. 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 non-redundant protein sequence entries in GenBank, but a reliably curated protein database such as Swiss-Prot 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.''

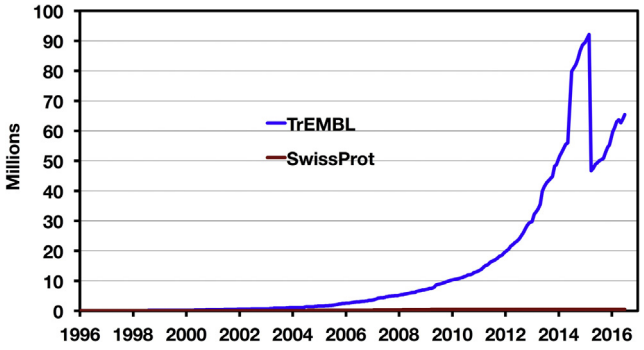

Figure 1 Growth of the UniProt database.

Sequence homology alone usually is used to assign the *in vitro* activities to uncharacterized enzymes; indeed, this is why 50% of the proteins in the sequence databases have incorrect or unknown functions (Schnoes et al., 2009). The sequence databases, UniProt and GenBank, use automated procedures to annotate the proteins discovered in genome projects (the TrEMBL database in UniProt) the function associated with the ''closest'' hit in the database is assigned as the function of the newly deposited protein; hence, the functions often are incorrect. To solve this problem, experimentalists require tools and strategies that facilitate informed mining of the sequence databases as well as facilitate access to complementary functional information, *e.g.*, genome context in the case of microbial proteins, thereby enabling the design of focused experiment-based assignment of *in vitro* activities and *in vivo* functions.

The Enzyme Function Initiative [EFI; formerly NIH U54GM093342 (5/01/10-4/30/15); now P01GM118303, 6/15/16-4/30/21 (Gerlt et al., 2011)] pioneered the development of multidisciplinary approaches to facilitate functional assignments; these included genomic enzymology/bioinformatics, protein production, structure determination, homology modeling, *in silico* ligand docking, experimental enzymology, microbiology, and metabolomics. In the EFI, the choice of targets for functional assignment was guided by ''genomic enzymology'', an ''expansive strategy for understanding the structural bases for catalysis''. As noted by Babbitt and Gerlt who popularized the term ''genomic enzymology'' (Gerlt and Babbitt, 2001): ''Until the early 1990s, enzymologists had little choice but to focus their studies on single examples of specific enzymes. Now, a much larger informational context is available, allowing enzymologists to include the genomic context (sequence families, structures, and functions) relevant to study of their favourite enzyme, rather than describing single-enzyme phenomenology.'' This larger context can be expected to facilitate the assignment of functions to uncharacterized proteins discovered in genome projects.

In particular, the EFI developed and popularized two large-scale ''genomic enzymology'' tools: (1) sequence similarity networks (SSNs) to allow analysis of sequence—function space in entire protein families, including the identification of isofunctional groups and the placement of restrictions on possible reactions and substrates and (2) genome neighbourhood networks (GNNs) to allow analysis of genome context (gene clusters and operons), thereby providing clues about the reactions, substrates, intermediates, and products in the metabolic pathways in which novel enzymes participate. These were used within the EFI to assign activities to novel enzymes in previously unknown metabolic pathways. As a result of these successes, the EFI made these ''genomic enzymology'' tools available to the community with ''user friendly'' web tools so that any enzymologist can mine the sequence databases for novel functions, *i.e.*, ''teaching a novice how to fish instead of serving fish''.

This article provides a brief description of the EFI's ''genomic enzymology'' web tools for generating SSNs and GNNs; it also provides an example of the use of these tools to discover and annotate novel enzymes in a novel metabolic pathway.

Sequence similarity networks, EFI-EST web tool

Dr. Patricia Babbitt, a member of the EFI until 2013, promoted the use of sequence similarity networks (SSNs) to allow large-scale visualization and analysis of sequence–function space in entire protein families (Atkinson et al., 2009). An SSN is the multidimensional homologue of a one-dimensional ''BLAST'' in which sequence relationships relating each member of a homologous protein family to all other members of the family are computed, visualized, and analysed. A symbol (''node'') in the SSN represents a sequence; lines (''edges'') connect sequence pairs to indicate relatedness (Fig. 2A). The numerical value of an edge (alignment score) is derived from the BLAST bit score from the sequence alignment. The power of SSNs is that as the alignment score threshold (sequence identity) for drawing edges is increased, the nodes segregate into clusters, allowing the user to assess the convergence/divergence of sequence and function as the sequences (nodes) segregate into putative isofunctional clusters.

SSNs are not as ''rigorous'' as trees and dendrograms in analyzing potential evolutionary relationships that require multiple sequence alignments that are more time-consuming to calculate than pairwise sequence identities (Fig. 2B); however, SSNs provide the advantage that they can be used interactively and, also, can provide the user with information about each protein that can be used to infer convergence of function, *e.g.*, phylogenetic relationships, membership in Pfam and InterPro families, and availability of three-dimensional structures. SSNs are visualized using Cytoscape, ''an open source software platform for visualizing complex networks and integrating these with attribute data''.

Dr. Babbitt is the driving force behind the Structure–Function Linkage Database (SFLD) (http://sfld.rbvi.ucsf.edu/django/) that provides the community with SSNs for a small number of functionally diverse enzyme superfamilies for which functional assignment is ''nontrivial'', *e.g.*, amidohydrolase, enolase, isoprenoid synthase, and radical SAM enzymes (Akiva et al., 2014). The functional annotations in these SSNs are reviewed and updated by curators, a labour-intensive/expensive process; therefore, the SFLD does not have the capability of providing the community with SNNs for *any* protein family.

Therefore, the EFI developed the EFI-Enzyme Similarity Tool (EFI-EST) to allow ''anyone'' to generate the SSN for their ''favorite'' protein family, with the node attribute information for the sequences mined primarily from the UniProtKB database ((http://www.uniprot.org/) (Gerlt et al., 2015). The UniProtKB database provides many types of ''bioinformatic'' information; the SwissProt database provides human-curated functional annotations mined from the literature. EFI-EST is available as a web tool (efi.igb.illinois.edu/efi-est/) that is maintained at the Institute for Genomic Biology, University of Illinois, Urbana-Champaign. The protein sequences used by EFI-EST are obtained from the UniProt database and are updated six times per year with each update of the InterPro database. The choice of the UniProt database instead of the GenBank database reflects the ability of the community to correct/update the annotations in the UniProt database; annotations in GenBank can be changed only by the depositor of the entry.

Generation of an SSN using the EFI-EST web tools involves three steps (Fig. 3):

(1) In the first step (''Start''), the user selects the method for collecting the sequences for the SSN:

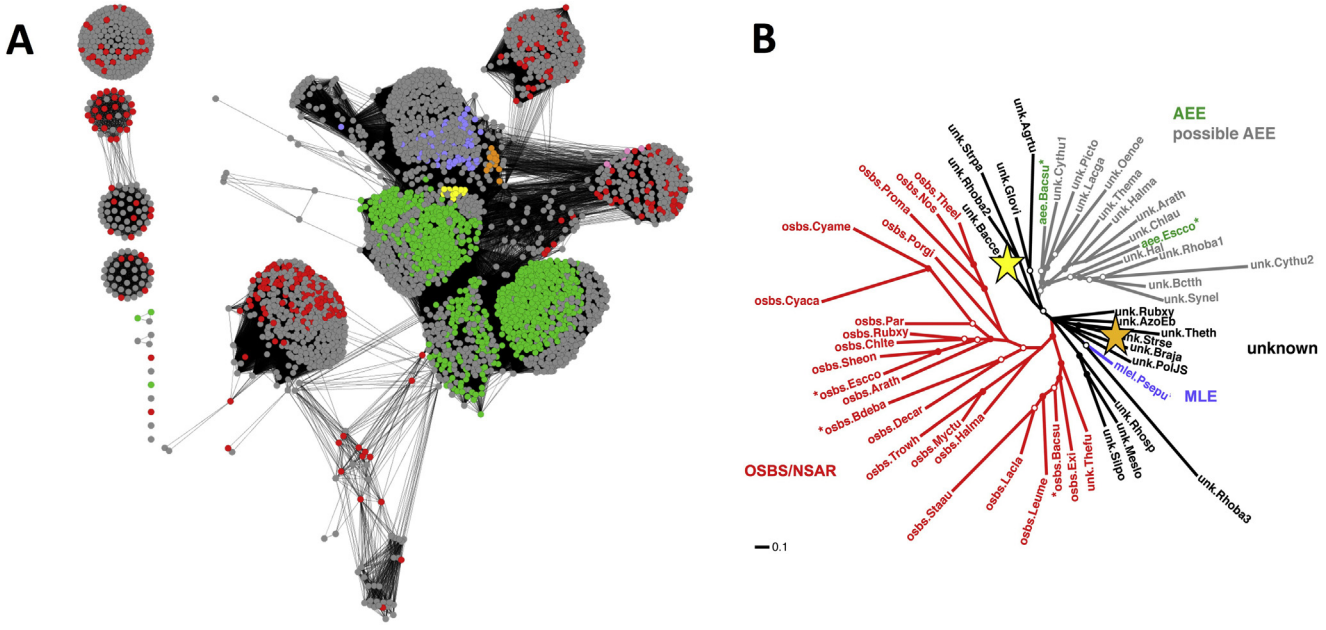

Figure 2 Comparison of a sequence similarity network (Panel A) and a dendrogram (Panel B) the muconate lactonizing enzyme subgroup of the enolase superfamily.

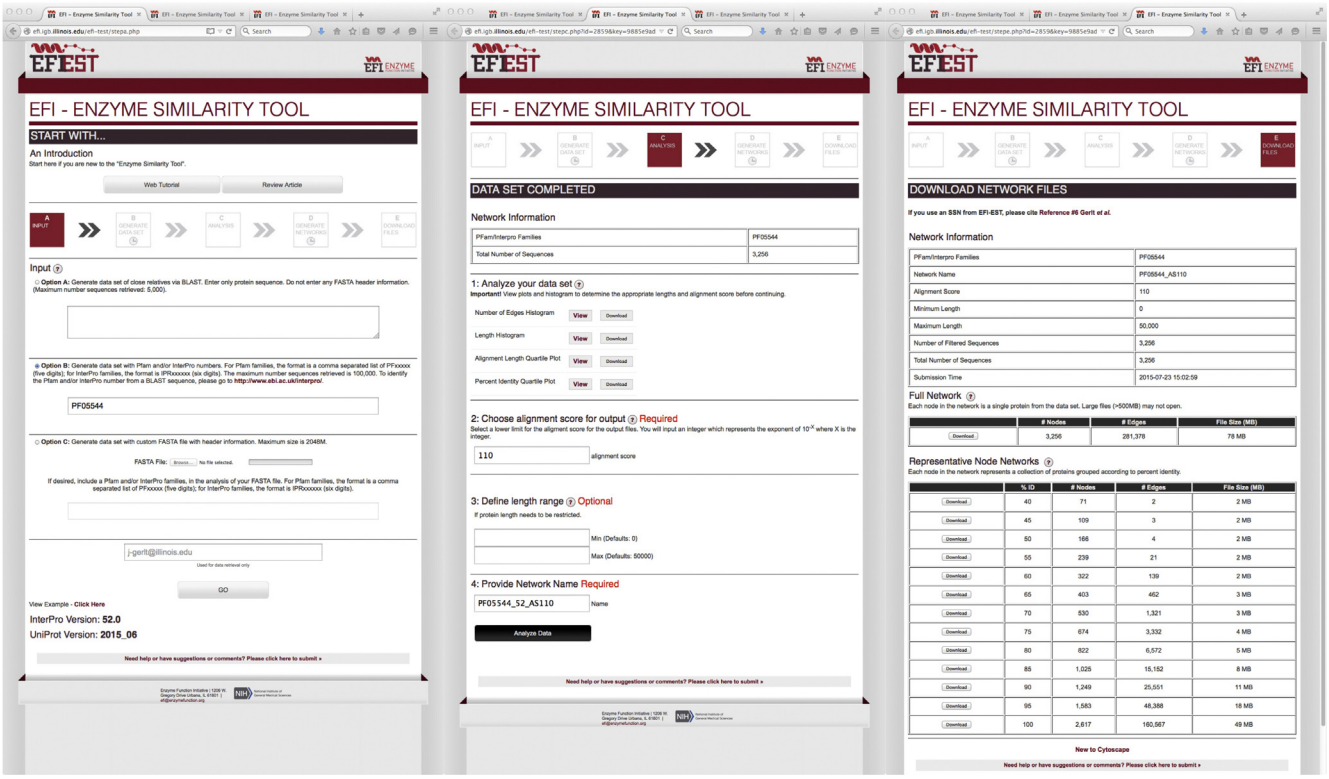

Figure 3 Screens for the step for generating a SSN with the EFI-EST web tool. Left, Start to input sequence/Pfam family. Centre, Data Set Completed/**Analyze** to input alignment score for SSNs. Right, **Download** Network Files to select/download SSN xgmml files.

Option A, a query sequence to collect the ''closest'' homologues using BLAST (default $\leq$5000 sequences);

Option B, any combination of Pfam (http://pfam.xfam.org/) and/or InterPro ((https://www.ebi.ac.uk/interpro/) families (default is the number of sequences in PF04055, the radical SAM superfamily; currently, $\leq$175,000 sequences), or

Option C, a FASTA file for upload that can be combined with any combination of Pfam and/or InterPro families.

(2) In the second step (''Analyze''), the user specifies a minimum alignment score (measure of sequence relatedness) for generating the SSN-sequence pairs with edges exceeding the alignment score are collected from the complete set generated by the all-by-all sequence comparison. The user is assisted in the selection of the alignment score by four histograms/quartile plots generated in the ''Start'' step, with the most essential providing the relationship between the alignment score and percent identity.

(3) In the third/final step (''Download''), the user downloads the SSN as one or more xgmml files. The user's RAM limits the number of edges that can be displayed by Cytoscape: with 4 GB RAM, an SSN with <500,000 edges can be opened; with 128 GB RAM, an SSN with <10,000,000 edges can be opened. Because the number of edges is unknown until after the all-by-all sequence comparison is performed, EFI-EST provides not only the full network (all sequences, if the network has $\leq$10M edges) but also representative node (rep node) networks. In rep node networks, sequences sharing greater than a specified percent identity are collected in the same meta-node, thereby reducing the number of nodes and edges. These meta-nodes contain sequences sharing from 40 to 100% sequence identity, in increments of 5% (for a total of 13 rep node networks). The user downloads the file(s) that his/her computer can open.

Genome neighbourhood networks, EFI-GNT web tool

In eubacteria, archaea, and fungi, the genes that encode metabolic pathways often are co-located in the genome in operons and gene clusters. Thus, the genome context of an uncharacterized enzyme can provide important clues about its *in vitro* enzymatic activity and *in vivo* physiological function by allowing identification of functionally related proteins that participate in metabolic pathways. To enable this, the EFI developed genome neighbourhood networks (GNNs) to enable large-scale visualization and analysis of genome context for one or more isofunctional clusters in the SSN for a protein family (Zhao et al., 2014).

The use of SSNs and GNNs is synergistic. With Cytoscape, the user first segregates the SSN for a protein family (from EFI-EST) containing the uncharacterized enzyme(s) into isofunctional clusters, *e.g.*, using known functions from SwissProt to choose an appropriate edge threshold. That SSN (or a subset of its clusters) is the input for the EFI-Genome Neighborhood Tool (EFI-GNT; http://efi.igb.illinois.edu/efi-gnt/) that interrogates the prokaryotic,

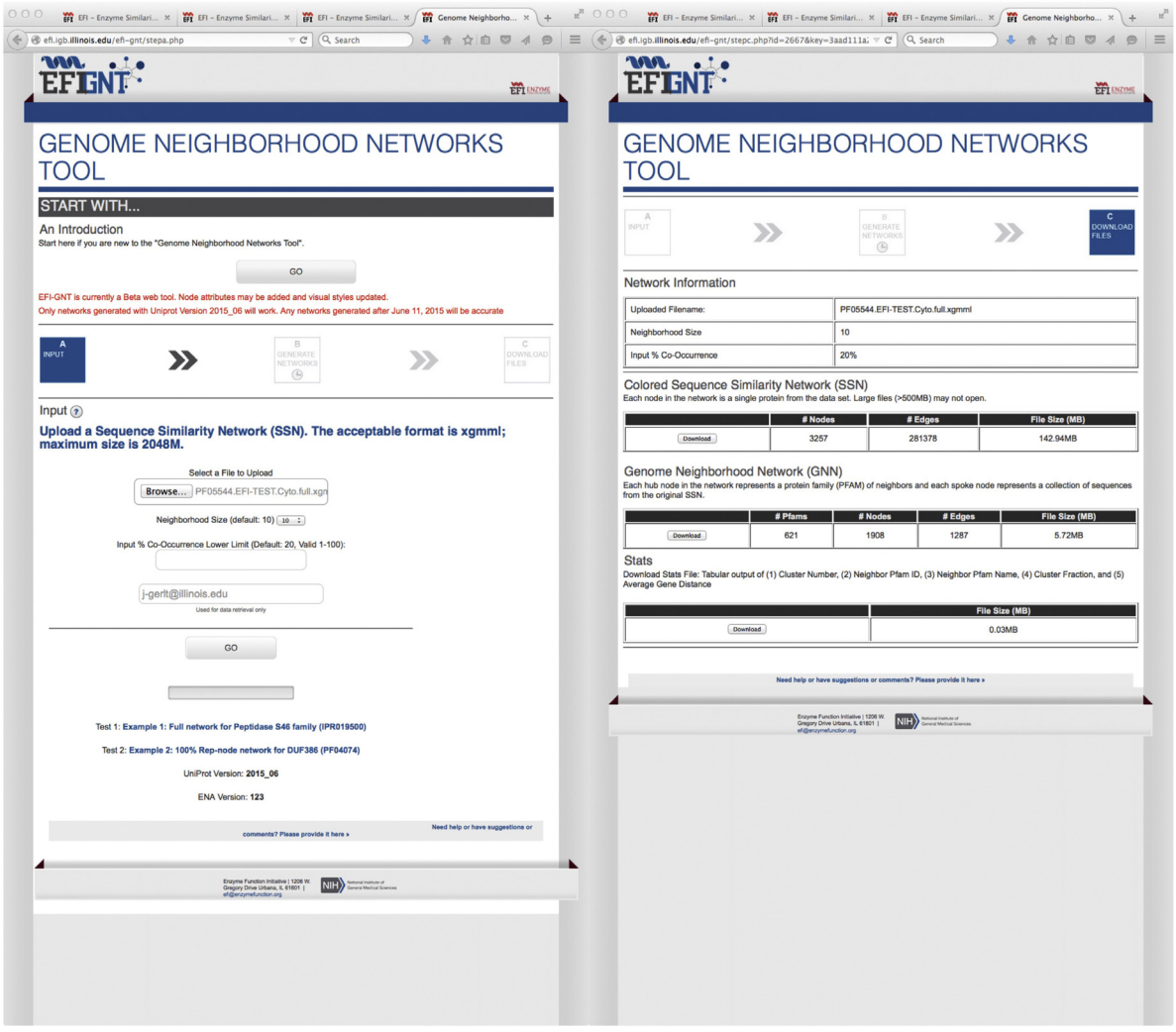

Figure 4 Screens for the step for generating a GNN with the EFI-GNT web tool. Left, **Start** screen to input SSN xgmml file. Right, **Download** Network Files screen to download SSN/GNN xgmml.

fungal, and metagenome ENA database for the genome neighbours of the query sequences within a user-specified gene window (default ± 10 genes). The neighbour proteins are associated with their Pfam families and displayed in the GNN (Fig. 3).

The GNN is used to (1) assess whether the edge threshold was appropriate to segregate the family into isofunctional clusters, *i.e.*, the genome contexts for the various clusters are distinct and (2) deduce the function/pathway of the clusters in the input SSN using the identities of the neighbor Pfam families and the locations of the neighbours in the SSNs for those families relative to known functions. For pathway discovery, the sequences in an ''isofunctional'' cluster in an SSN often are encoded by diverse species; because genome neighbourhoods often are not conserved phylogenetically, a GNN allows identification of pathway components that are not proximal in the genome of the organism that encodes the target uncharacterized enzyme. With the identities of the Pfam families, the user can infer the types of reactions in the metabolic pathway. With experimental information about the identity of the substrate for the first enzymes in the pathway, the user can infer the substrate, intermediates, product, and reactions in the pathway (*vide infra*).

Generation of a GNN using the EFI-GNT web tool involves two steps (Fig. 4):

(1) In the first step (''Start''), the user uploads the xgmml file for one or more clusters from an SSN generated with EFI-EST or filtered/output from Cytoscape. EFI-GNT then collects the proteins encoded by genes proximal to those that encode the query sequences in the input SSN from the ENA database (prokaryote, archaeal, fungal, and metagenome sequences that provide functionally relevant genome context because pathways in these organisms often are encoded by operons and/or gene clusters). The user can specify the size of the gene ''window'' surrounding the query (default ± 10, the ''signal to noise'' for functionally linked neighbours often can be increased by decreasing the window size; the optimum size depends on the complexity of the operon/gene cluster than encodes the components of the pathway). The neighbours are collected into Pfam

families (~80% of the sequences in the UniProt database are associated with one or more Pfam families) for inclusion in the GNN. The sequences not associated with a Pfam family are provided in the GNN because their functions may be discovered as new pathways are annotated.

(2) In the second step (''Download''), the user downloads the GNN. In the current configuration of EFI-GNT, the neighbor Pfam families are the ''hub'' nodes (Fig. 6A); the neighbours identified by the sequences in the query clusters are located in ''spoke'' nodes for each query cluster. This presentation allows the user to assess whether multiple clusters in the input SSN share genome neighbours with the same function (same Pfam family, assuming that the genome proximal members of the Pfam family are orthologues). The user can determine whether the input SSN was segregated into isofunctional clusters or ''over-fractionated'', *i.e.*, multiple clusters have the same genome contexts because the user separated (accidentally or intentionally) query orthologues into multiple clusters because of either conserved functions in divergent clusters or phylogenetic divergence.

Transport system solute binding protein (sbp)-guided pathway discovery

The considerable use of the synergistic use of SSNs and GNNs to facilitate the assignment of novel *in vitro* activities and *in vivo* metabolic functions to uncharacterized enzymes discovered in genome projects is best established with an example.

The EFI developed an experimental platform to achieve this goal by exploiting its ability for large-scale protein production and ligand screening to discover the ligand specificities of the extracellular/periplasmic solute binding proteins (SBPs) for microbial ABC, TRAP, and TCT transport systems (Vetting et al., 2015). Because the genes that encode transport systems often are co-localized on the genome with the genes that encode the catabolic pathway for the transported solute, the specificities of the SBPs can be used to identify the substrate for the first enzyme in the catabolic pathway and its genome neighbours will identify the enzymes in the catabolic pathway. With the Pfam family membership of the pathway enzymes, *e.g.*, aldolase, oxidase, transaminase, or kinase, the catabolic pathway for the SBP ligand can be inferred.

The EFI first focused on the SBPs for the TRAP (**TR**ipartite **A**TP-independent **P**eriplasmic) transporters. Although not as ubiquitous as ABC (**A**TP **B**inding **C**assette) transporters, we observed that the genes encoding TRAP transporters frequently are colocated with genes encoding catabolic pathways for acid sugars (substrates for members of the enolase superfamily, one of the functionally diverse superfamilies that were explored by the EFI). More than 300 TRAP SBPs were placed in the EFI's protein production pipeline; 158 were purified after heterologous expression in *Escherichia coli*. Differential scanning fluorimetry (DSF, aka Thermofluor) was used to screen the purified SBPs for ''hits'' using a library of 189 small-molecule ligands that was enriched with carbohydrate derivatives (including all D- and L-hexoses, pentoses, and tetroses, their aldonic and aldaric acid derivatives, and all 16 D- and L-hexuronic acids) but also included D- and L-amino acids, aromatic acids, and known TRAP ligands (24 known at the time of the study in 2015). Eight-nine purified SBPs yielded a DSF ''hit'', with virtually all of these carboxylate-containing ligands as was expected based on the known ligands. Structures were determined for many of these SBPs in the presence of their ligands, providing a valuable database of experimental structures to guide homology modelling and virtual docking to identify ligands for SBPs that had not been purified or for prediction of ligands using a larger virtual ligand library.

When possible, structures were also determined for ''apo'' SBPs. Interestingly and importantly, several of these structures revealed the presence of tightly bound ligands from the *E. coli* metabolome that co-purified with the SBP. These included orotic acid, glycerol phosphate, diglycerol phosphate, indole acetate, and ethanolamine that were not present in the DSF ligand library. The co-purification of these metabolites suggests that these are *in vivo* ligands for SBPs, thereby allowing these to be used for catabolic by the encoding organism.

The SSN for the TRAP SBP family (Pfam family PF03480) was constructed, and the DSF ''hits'' were mapped to the SSN. The before/after annotated SSN comparison in Fig. 5 demonstrates the considerable power of this large-scale screening approach for exploring ligand specificity (function)-sequence space in the family. The SSNs in the Figure were constructed using an alignment score of 120 that corresponds to ~60% sequence identity.

As described in the next section, we selected the cluster from this SSN that contained the SBP that co-purified with ethanolamine (from *Chromohalobacter salexigens*) for discovery of a novel catabolic pathway for ethanolamine.

Discovery of a novel pathway for ethanolamine catabolism

The cluster containing the ethanolamine-binding SBP from *C. salexigens* (68 sequences) was used as the input for the EFI-GNT web tool so that the pathway for catabolism of ethanolamine could be identified. The total GNN contains, as expected, the Pfam families for the membrane components of the TRAP transport system, a variety of transcriptional regulators, and several candidates for the catabolic enzymes.

The enzyme families that frequently co-occur with the ethanolamine-binding SBPs (Fig. 6) include the glutamine synthase family (Gln-synt_C; PF00120), the iron-dependent alcohol dehydrogenase family (Fe-ADH; PF00465), the aldehyde dehydrogenase family (Aldeh; PF00171), and the formylglutamate amidohydrolase family (FGase; PF05013). A previous study had identified a pathway for isopropylamine catabolism (to L-alanine) that was initiated by the ATP-dependent conjugation of isopropylamine with L-glutamate to form γ-glutamylisopropylamide (Fig. 7A; de Azevedo Wäsch et al., 2002)). In analogy with this pathway and using the Pfam families identified by the GNN as a guide, we postulated a previously unknown pathway for ethanolamine catabolism (Fig. 7B). In this pathway, γ-glutamylethanolamide is generated by the member of the glutamine synthase family (PF00120; ethanolamine γ-glutamylase); the hydroxyl group is successively oxidized to

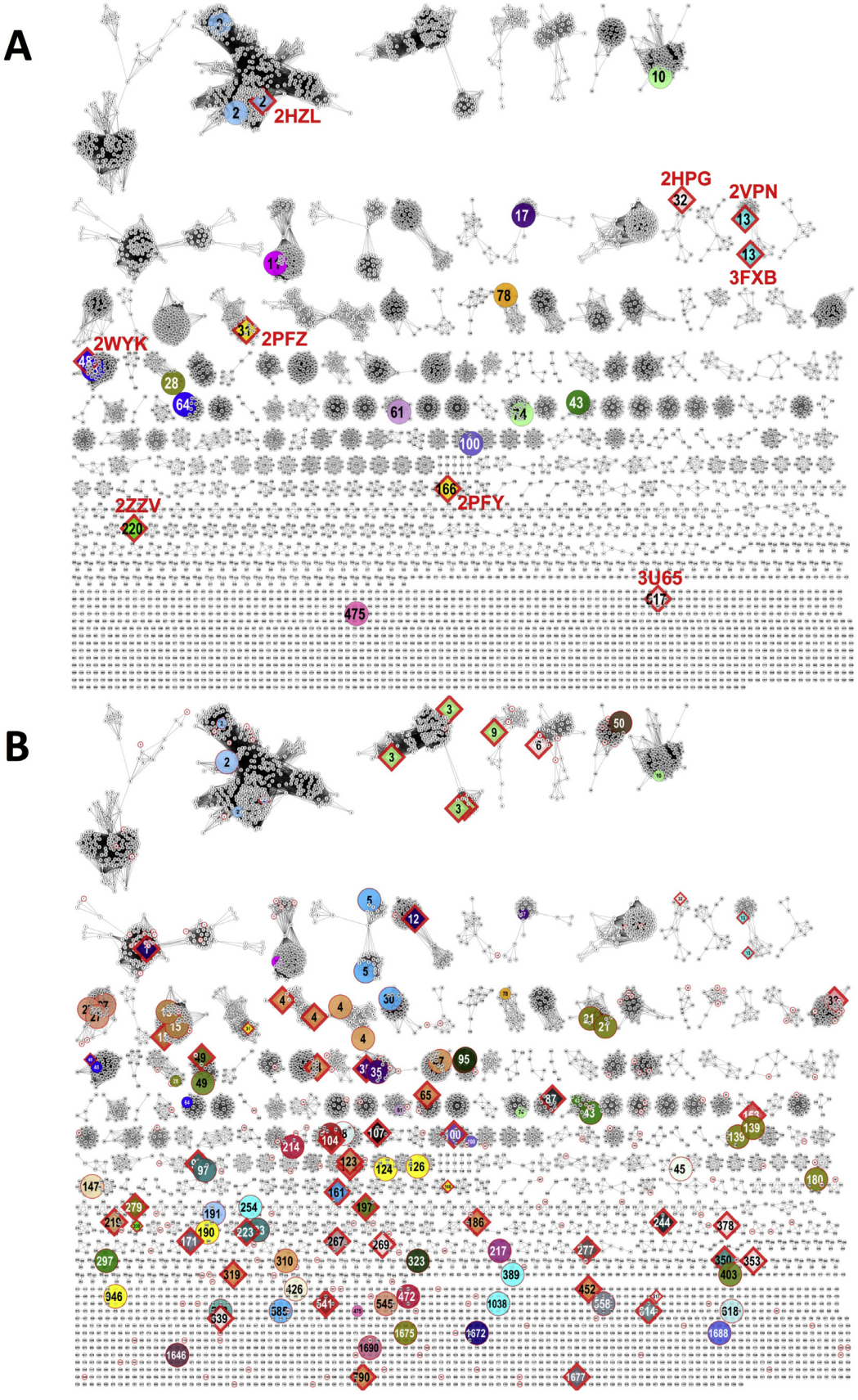

Figure 5 Comparison of the ligand-specificity annotated SSNs for the TRAP SBP family (PF03480) before (Panel A) and after (Panel B) the EFI's large-scale protein production/ligand screening project.

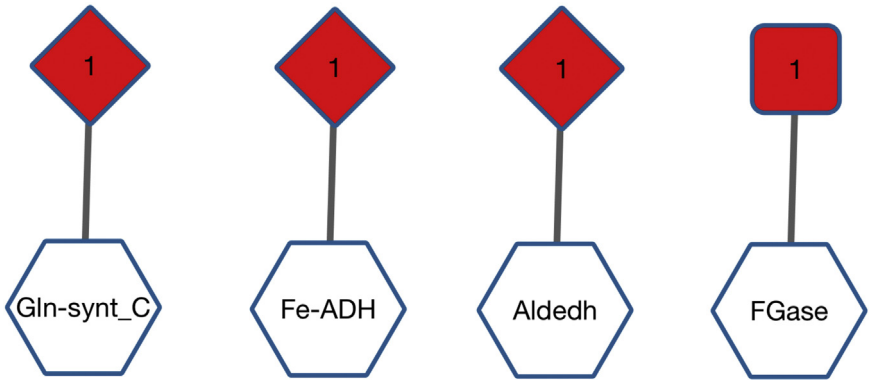

Figure 6 Selected clusters from the GNN generated with the cluster of ethanolamine-binding SBPs showing the four enzyme families involved in the ethanolamine catabolic pathway in *Chromohalobacter salexigens*.

an aldehyde and then a carboxylate group by members of the iron-dependent alcohol dehydrogenase family (PF00465) and the aldehyde dehydrogenase family (PF00171), respectively. The resulting γ-glutamylglycine is hydrolysed to glycine and L-glutamate (the latter a ''catalyst'' in the pathway) by a member of the formylglutamate amidohydrolase family (PF05013). The ubiquitous glycine cleavage pathway oxidizes glycine to ammonia, 5,10-methylenetetrahydrofolate, and CO_2, thereby providing the encoding organism with the ability to utilize ethanolamine as sole nitrogen source.

We determined that *C. salexigens* utilizes ethanolamine as sole nitrogen source. We also constructed knockouts of the genes encoding the SBP and the ethanolamine γ-glutamylase — both were unable to utilize ethanolamine as sole nitrogen source. As further evidence for this pathway, we determined that the genes for the TRAP transporter (SBP and two membrane components) as well as all four catabolic enzymes are upregulated in the presence of ethanolamine. We also purified the ethanolamine γ-glutamylase and γ-glutamylglycine hydrolase (PF05013) and determined their kinetic constants; these were consistent with the *in vitro* activities deduced from their Pfam membership. And, finally, we performed metabolomic studies that detected the presence of the predicted γ-glutamylethanolamide, γ-glutamylaminoacetaldehyde, and γ-glutamylglycine when *C. salexigens* was grown on ethanolamine as sole nitrogen source. Thus, this pathway for ethanolamine catabolism is secure.

Prior to these studies, a pathway for ethanolamine catabolism (sole carbon source) had been identified

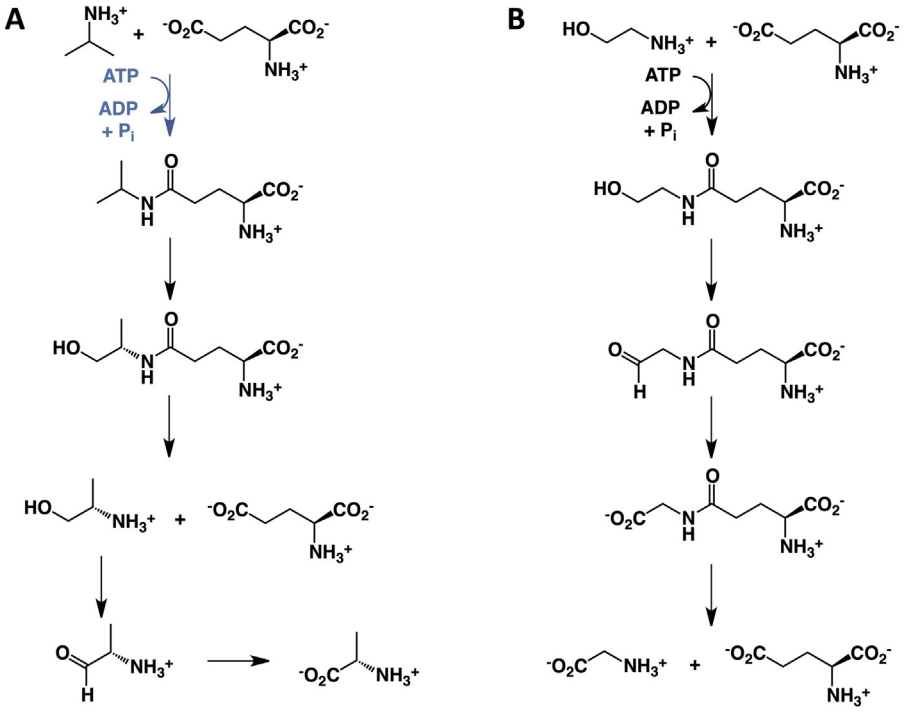

Figure 7 Panel A, pathway for isopropylamine catabolism initiated by an isopropylamine γ-glutamylase. Panel B, pathway for ethanolamine catabolism initiated by an ethanolamine γ-glutamylase.

that uses the adenosylcobalamin-dependent ethanolamine ammonia lyase that produces acetaldehyde and ammonia (Garsin, 2010). The acetaldehyde then can be converted to acetyl-CoA, acetyl phosphate, and acetate, providing intermediates in known catabolic pathways.

Additional examples of SBP-guided catabolic pathway discovery

We have used the ''same'' SBP-guided strategy (using additional clusters in the TRAP SBP family as well as clusters in the SSNs for ABC SBP families) to discover additional novel pathways (Wichelecki et al., 2015; Huang et al., 2015). Of particular note, we used this approach to assign novel kinase functions to many members of a **D**omain of **U**nknown **F**unction (DUF1537, PF07005) (Zhang et al., 2016). The power lies in the large-scale screening of SBPs with a ligand library to identify novel candidates for previously unknown catabolic pathways. Approximately 20% of the protein families curated by Pfam are members of DUFs, highlighting the expected occurrence of novel enzymes in novel metabolic pathways amongst the uncharacterized enzymes discovered in genome projects.

Summary

As described in this brief review, the EFI developed ''genomic enzymology'' approaches, including publicly accessible web tools, to enable the community to mine the sequence databases for novel enzymes in novel metabolic pathways. The annotation challenge is too large for a single project to solve. However the development of ''user friendly'' tools and strategies that can be adopted by the enzymology, chemical biology, systems biology, and microbiology communities has the potential to improve the quality of annotations in the sequence databases, thereby making them more useful and valuable as genome projects continue.

References

Akiva, E., Brown, S., Almonacid, D.E., Barber II, A.E., Custer, A.F., Hicks, M.A., Huang, C.C., Lauck, F., Mashiyama, S.T., Meng, E.C., Mischel, D., Morris, J.H., Ojha, S., Schnoes, A.M., Stryke, D., Yunes, J.M., Ferrin, T.E., Holliday, G.L., Babbitt, P.C., 2014. The Structure—Function Linkage Database. Nucleic Acids Res. 42, D521—D530, http://dx.doi.org/10.1093/nar/gkt1130.

Atkinson, H.J., Morris, J.H., Ferrin, T.E., Babbitt, P.C., 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One 4, e4345, http://dx.doi.org/10.1371/journal.pone.0004345.

de Azevedo Wäsch, S.I., van der Ploeg, J.R., Maire, T., Lebreton, A., Kiener, A., Leisinger, T., 2002. Transformation of isopropylamine to L-alaninol by *Pseudomonas* sp. strain KIE171 involves *N*-glutamylated intermediates. Appl. Environ. Microbiol. 68, 2368—2375, http://dx.doi.org/10.1128/aem.68.5.2368-2375.2002.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Dougherty, B.A., Merrick, J.M., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496—512, http://dx.doi.org/10.1126/science.7542800.

Garsin, D.A., 2010. Ethanolamine utilization in bacterial pathogens: roles and regulation. Nat. Rev. Microbiol. 8, 290—295, http://dx.doi.org/10.1038/nrmicro2334.

Gerlt, J.A., Babbitt, P.C., 2001. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu. Rev. Biochem. 70, 209—246, http://dx.doi.org/10.1146/annurev.biochem.70.1.209.

Gerlt, J.A., Allen, K.S., 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.S., Shoichet, B.K., Sweedler, J.V., 2011. The enzyme function initiative. Biochemistry 50, 9950—9962, http://dx.doi.org/10.1021/bi201312u.

Gerlt, J.A., Bouvier, J.A., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., Whalen, K.L., 2015. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): a web tool for generating protein sequence similarity networks. Biochim. Biophys. Acta 1854, 1019—1037, http://dx.doi.org/10.1016/j.bbapap.2015.04.015.

Huang, H., Carter, M.S., Vetting, M.W., Al-Obaidi, N., Patskowsky, Y., Almo, S.C., 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, http://dx.doi.org/10.1021/jacs.5b08968.

Khosla, C., 2015. Quo vadis, enzymology? Nat. Chem. Biol. 11, 438—441, http://dx.doi.org/10.1038/nchembio.1844.

Schnoes, A.M., Brown, S.D., Dodevski, I., Babbitt, P.C., 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. 5, e1000605, http://dx.doi.org/10.1371/journal.pcbi.1000605.

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., Rodinov, 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., 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, http://dx.doi.org/10.1021/bi501388y.

Wichelecki, D.J., Vetting, M.W., Chou, L., Al-Obaidi, N., Bouvier, J.T., Almo, S.C., 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, http://dx.doi.org/10.1074/jbc.m115.686857.

Zhang, X., Carter, M.S., Vetting, M.W., San Francisco, B., Zhao, S., Al-Obaidi, N.F., Solbiati, J.O., Thiaville, J.J., de Crécy-Lagard, V., 2016. Assignment of function to a domain of unknown function (DUF): DUF1537 is a new kinase family in catabolic pathways for acid sugars. Proc. Natl. Acad. Sci. U. S. A., http://dx.doi.org/10.1073/pnas.1605546113.

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., Jacobson, M.P., 2014. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. eLife 3, e03275, http://dx.doi.org/10.7554/elife.03275.