HHS Public Access

Author manuscript

J Ind Microbiol Biotechnol. Author manuscript; available in PMC 2017 March 01.

Published in final edited form as:

J Ind Microbiol Biotechnol. 2016 March ; 43(0): 261–276. doi:10.1007/s10295-015-1671-0.

Genome Neighborhood Network Reveals Insights into Enediyne Biosynthesis and Facilitates Prediction and Prioritization for Discovery

Jeffrey D. Rudolf[1,‡], **Xiaohui Yan**[1,‡], and **Ben Shen**[1,2,3,*]

[1]Department of Chemistry, The Scripps Research Institute, Jupiter, FL 33458, USA

[2]Department of Molecular Therapeutics, The Scripps Research Institute, Jupiter, FL 33458, USA

[3]Natural Products Library Initiative, The Scripps Research Institute, Jupiter, FL 33458, USA

Abstract

[*]Corresponding author: Ben Shen; tel.: +1 (561) 228-2456; fax: +1 (561) 228 2472; shenb@scripps.edu.
[‡]J. D. R. and X. Y. contributed equally to this work.

GenBank accession numbers

Published enediyne gene clusters or PKS cassettes with GenBank accession numbers (in parentheses): *Actinomadura madurae* ATCC 39144 (MDP, AY271660), *Actinomadura verrucosospora* ATCC 39334 (ESP, AY267372), *Micromonospora chersina* ATCC 53710 (DYN, EF552206), *Micromonospora echinospora* NRRL 15839 (CAL, AF497482), *Salinispora pacifica* CNS143 (CYA, KC863955), *Salinispora tropica* CNB-440 (SPO, CP000667), *Streptoalloteichus* sp. ATCC 53650 (KED, JX679499), *Streptomyces carzinostaticus* subsp. neocarzinostaticus ATCC 15944 (NCS, AY117439), *Streptomyces globisporus* C-1027 (C-1027, AY048670), *Streptomyces* sp. CNT179 (CYN, KC863954). Putative enediyne producers with GenBank accession numbers for their respective genomes (in parentheses): *Actinoalloteichus cyanogriseus* DSM 43889 (AUBJ00000000), *Actinokineospora inagensis* DSM 44258 (AXWW00000000), *Actinomadura flavalba* DSM 45200 (ARFO00000000), *Actinomadura oligospora* ATCC 43269 (JADG00000000), *Actinoplanes* sp. N902-109 (CP005929), *Actinopolyspora halophila* DSM 43834 (AQUI00000000), *Actinospica robiniae* DSM 44927 (AZAN00000000), *Amycolatopsis alba* DSM 44262 (ARAF00000000), *Amycolatopsis azurea* DSM 43854 (ANMG00000000), *Amycolatopsis balhimycina* FH 1894 DSM 44591 (ARBH00000000), *Amycolatopsis decaplanina* DSM 44594 (AOHO00000000), *Amycolatopsis halophila* YIM 93223 (AZAK00000000), *Amycolatopsis japonica* MG417-CF17 (CP008953), *Amycolatopsis mediterranei* DSM 40773 (JMQJ00000000), *Amycolatopsis mediterranei* RB (CP003777), *Amycolatopsis nigrescens* CSC17Ta-90 DSM 44992 (ARVW00000000), *Amycolatopsis orientalis* HCCB 10007 (CP003410), *Amycolatopsis rifamycinica* 46095 (JMQI00000000), *Bacillus* sp. 2_A_57_CT2 (ACWD00000000), *Catenulispora acidiphila* DSM 44928 (CP001700), *Cyanothece* sp. PCC 7822 (CP002198), *Haliangium ochraceum* DSM 14365 (CP001804), *Herpetosiphon aurantiacus* DSM 785 (CP000875), *Kitasatospora cheerisanensis* KCTC 2395 (JNBY00000000), *Methylomicrobium alcaliphilum* (NC_016112), *Methylomicrobium buryatense* 5G (AOTL00000000), *Microcystis aeruginosa* NIES-843 (AP009552), *Micromonospora aurantiaca* ATCC 27029 (CP002162), *Micromonospora lupini* str. Lupac 08 (CAIE00000000), *Micromonospora* sp. CNB394 (ARGW00000000), *Micromonospora* sp. L5 (CP002399), *Nocardia brasiliensis* ATCC 700358 (CP003876), *Nocardia brasiliensis* IFM 10847 (BAUA00000000), *Nocardia brasiliensis* NBRC 14402 (BAFT00000000), *Nocardia tenerifensis* NBRC 101015 (BAGH00000000), *Nocardia* sp. CNY236 (AXVD00000000), *Nocardiopsis ganjiahuensis* DSM 45031 (ANBA00000000), *Nocardiopsis* sp. CNT312 (AZXF00000000), *Nostoc* sp. PCC 7524 (CP003552), *Saccharomonospora halophila* 8 (AICX00000000), *Saccharomonospora marina* XMU15 (CM001439), *Saccharomonospora* sp. CNQ490 (AZUM00000000), *Saccharothrix espanaensis* DSM 44229 (HE804045), *Salinispora arenicola* CNS-205 (CP000850), *Salinispora arenicola* CNS-991 (ABB00000000), *Salinispora pacifica* CNS237 (AUGH00000000), *Salinispora pacifica* DSM 45549 (AQZW00000000), *Salinispora tropica* CNR699 (ARHP00000000), *Scytonema hofmanni* PCC 7110 (ANNX00000000), *Streptomyces clavuligerus* ATCC 27064 plasmid pSCL4 (CM000914), *Streptomyces griseus* subsp. *griseus* NBRC 13350 (AP009493), *Streptomyces griseus* XyelbKG-1 (ADFC00000000), *Streptomyces* sp. ATexAB-D23 (AREH00000000), *Streptomyces* sp. CNR698 (AZXC00000000), *Streptomyces* sp. CNS615 (AQPE00000000), *Streptomyces* sp. CNT302 (ARIM00000000), *Streptomyces* sp. LaPpAH-95 (AQWQ00000000), *Streptomyces* sp. LaPpAH-165 (AREB00000000), *Streptomyces* sp. SA3_actG (ADXA00000000), *Streptomyces* sp. Tu6071 (CM001165), *Streptomyces* sp. W007 (AGSW00000000), *Streptomyces viridosporus* T7A ATCC 39115 (AJFD00000000), *Streptomyces yeochonensis* CN732 (JQNR00000000), *Streptosporangium roseum* DSM 43021 (CP001814), *Verrucosispora maris* AB-18-032 (CP002638), *Tolypothrix* sp. PCC 7601 UTEX B 481 (AGCR00000000).

JGI/IMG accession numbers

Redundant genomes found in both databases are listed with only GenBank accession numbers. Putative enediyne producers listed here with JGI/IMG accession for their respective genomes (in parentheses) were not found in GenBank. *Calothrix desertica* PCC 7102 (2509887024), *Kitasatospora* sp. SolWspMP-SS2h (2524614568).

The enediynes are one of the most fascinating families of bacterial natural products given their unprecedented molecular architecture and extraordinary cytotoxicity. Enediynes are rare with only 11 structurally characterized members and four additional members isolated in their cycloaromatized form. Recent advances in DNA sequencing have resulted in an explosion of microbial genomes. A virtual survey of the GenBank and JGI genome databases revealed 87 enediyne biosynthetic gene clusters from 78 bacteria strains, implying enediynes are more common than previously thought. Here we report the construction and analysis of an enediyne genome neighborhood network (GNN) as a high-throughput approach to analyze secondary metabolite gene clusters. Analysis of the enediyne GNN facilitated rapid gene cluster annotation, revealed genetic trends in enediyne biosynthetic gene clusters resulting in a simple prediction scheme to determine 9- vs 10-membered enediyne gene clusters, and supported a genomic-based strain prioritization method for enediyne discovery.

Keywords

Enediyne polyketide synthase; Genome neighborhood network; Biosynthetic gene cluster; Genome mining; Natural products

INTRODUCTION

The enediyne natural products possess unprecedented molecular architecture and extraordinary cytotoxicity making them one of the most fascinating families of natural products known to date. All enediynes are comprised of an unsaturated core containing two acetylenic groups conjugated to a double bond or incipient double bond [13,27]. This conserved enediyne core is responsible for DNA damage and ultimately, cell death [10,21]. After the enediyne binds to DNA, the enediyne core, or warhead, undergoes electronic cyclization via a Bergman or Myers-Saito rearrangement affording a benzenoid diradical, which abstracts hydrogen atoms from the deoxyribose backbone of DNA [17,20,35]. Due to their extreme potencies (IC_{50}s against selected cancer cell lines are 10 pM – 10^{-3} pM) [47,58] enediynes must be harnessed and delivered to target cells for use as effective anticancer drugs [3,43]. Indeed, various polymer-based delivery systems or antibody-drug conjugates (ADCs) using enediynes as payloads have been developed and show clinical success or promise [7,43,46,50].

The enediyne family of natural products are rare with only 11 structurally characterized members and four additional members isolated in their cycloaromatized form (Fig. 1). Enediynes are separated into two subclasses based on the number of carbons in their core ring: 9-membered, which are commonly associated with a protective, sequestration apoprotein, and 10-membered [13,27]. The 9-membered enediynes consist of neocarzinostatin (NCS) from *Streptomyces carzinostaticus* [11], C-1027 from *S. globisporus* [40]. kedarcidin (KED) from a *Streptoalloteichus* sp. [41], maduropeptin (MDP) from *Actinomadura madurae* [23], and N1999A2 from *Streptomyces* sp. AJ9493 [22]. The 9-membered sporolides (SPOs) from *Salinispora tropica* CNB-440 [5], cyanosporasides (CYAs) from *Salinispora pacifica* CNS-143 [38], cyanosporasides (CYNs) from *Streptomyces* sp. CNT-179 [25], and fijiolides from *Nocardiopsis* sp. CNS-653 [37], were

isolated in the cycloaromatized form. The discrete 10-membered enediynes consist of calicheamicin (CAL) from *Micromonospora echinospora* [26], esperamicin (ESP) from *Actinomadura verrucosospora* [15], dynemicin (DYN) from *M. chersina* [36], uncialamycin (UCM) from *S. uncialis* [9], and namenamicin and the shishijimicins from the marine ascidia *Polysyncraton lithostrotum* and *Didemnum proliferum*, respectively [33,39].

The recent advances in DNA sequencing have resulted in an explosion of microbial genomes and gene clusters available in the public databases. As of May 13, 2015, there were 6311 and 25906 bacterial genomes in the National Center for Biotechnology Information (NCBI) and Joint Genome Institute (JGI) online databases, respectively. Analysis of these microbial genomes illustrates that the biosynthetic potential of bacteria, especially Actinobacteria (921 and 3035 in GenBank and JGI, respectively), is greatly underappreciated. The enediynes are no exception: a recent virtual survey revealed a total of 61 enediyne biosynthetic gene clusters, 54 of which are from the order Actinomycetales [45]. Ten of these gene clusters (NCS [31], C-1027 [30], KED [32], MDP [52], SPO [34], CYA [25], CYN [25], CAL [1], ESP [partial] [29,54], and DYN [14]) are responsible for the biosynthesis of known enediynes.

Remarkably, after almost 15 years of studying 9- and 10-membered enediyne gene clusters (the gene clusters of both C-1027 and CAL were reported in 2002) [1,30], there are still many unknowns regarding enediyne biosynthesis, regulation, and resistance. While significant progress has been made towards elucidating the biosynthesis of the peripheral moieties present in enediynes [27,53], the biosynthesis of the enediyne core has yet to be revealed. Comparison of the known enediyne gene clusters revealed a conserved set of five genes named the enediyne polyketide synthase (PKS) cassette [45]. Each cassette contains genes that encode an enediyne PKS (PKSE), a thioesterase (E10 or E7), and three unknown proteins (E3, E4, and E5 or U15, U14, and T3, respectively). This five gene cassette forms an apparent operon and is commonly arranged as *E3/E4/E5/pksE/E10* (Fig. 2A). The PKSE, an iterative type I PKS, initiates both 9- and 10-membered enediyne core biosynthesis through the production of a linear polyene intermediate [2,18,24,55]. However, the enzymes and chemistry responsible for the transformation of the polyene, or the corresponding ACP-bound intermediate, into the 9- and 10-membered enediyne cores are still unknown. The differences between the carbocycles of 9- and 10-membered enediynes suggest specific associated enzymes exist for both pathways (Fig. 2B). However, the lack of fully sequenced, 10-membered enediyne gene clusters (only CAL and DYN are available) has severely hampered the ability to identify genes conserved in 10-membered, and absent in 9-membered, enediyne gene clusters. Due to these reasons, the ability to accurately predict 9- vs 10-membered enediynes based solely on genetic parameters is currently lacking. Increasing the population of 10-membered enediyne gene clusters using available genomes will facilitate the discovery of bioinformatics trends.

The discovery of novel natural products found within genetically amenable and high producing microbial strains is always a desired and ideal research objective. However, even the rediscovery of natural products from alternative strains with improved characteristics, such as higher titers and/or genetic amenability, can alleviate technical difficulties encountered with “problem” strains [42]. Enediyne producers have been notoriously difficult

to work with. The limited number of enediyne producers, the lack of genetic amenability for many of these strains, the inherent instability of enediynes, and low production titers of the natural products, congeners, or intermediates, have all impeded their biosynthetic study and development as a drug lead. Therefore, the ability to effectively prioritize strains and gene clusters of interest, whether it be for the discovery of known or novel enediynes, is an urgent need for continued success in the enediyne field.

Recently, genome neighborhood networks (GNNs) were described as a bioinformatics strategy to predict enzymatic functions on a large scale based on their genomic context [57]. Construction and analysis of a proline racemase (PR) superfamily GNN by including ±10 genes in relation to the gene encoding each PR facilitated accurate predictions, which were subsequently experimentally verified, of many PR enzymes including new members of the superfamily [57]. The clustered nature of genes responsible for the production of bacterial natural products suggests GNNs could be a valuable tool for gene cluster annotation and natural product discovery. Given the size and complexity of the known enediyne gene clusters (up to 100 kb), the large number of functionally unassigned proteins in these gene clusters, and a desire to incorporate the numerous putative enediyne gene clusters, we considered an enediyne GNN as a viable way to quickly and accurately analyze these complicated natural product gene clusters.

Here we report the construction and analysis of an enediyne GNN. A virtual survey of the GenBank and JGI genome databases resulted in 87 potential enediyne gene clusters from 78 different bacteria strains, supporting Actinomycetales as the most prolific enediyne producers. Utilizing these bacterial genomes, an enediyne GNN was constructed revealing (i) the effectiveness of GNNs to probe natural product biosynthesis in a high-throughput manner, (ii) a visual way to quickly assign proteins with enediyne functionality and essentiality while simultaneously eliminating non-enediyne “noise”, (iii) a simple prediction scheme to determine 9- vs 10-membered enediyne gene clusters, and (iv) a genomic-based strain prioritization method for enediyne discovery. This study supports the development and use of bioinformatics tools during the discovery of natural products, including the continued pursuit for novel enediynes and alternative enediyne producers, in the genomic era.

RESULTS AND DISCUSSION

Virtual survey of genome databases highlighting the underexplored nature of enediynes

The unique *E3/E4/E5/pksE/E10* organization of the *pksE* cassette (Fig. 2A) provides a viable search query to mine enediyne gene clusters from the public databases. Homology searches using the proteins found within the C-1027 PKSE cassette, SgcE, SgcE3, SgcE4, SgcE5, and SgcE10, as queries in the nonredundant (NR) and JGI/IMG genome databases afforded 271 total enediyne gene clusters from 192 different bacteria strains encompassing 29 different genera. Gene clusters were proposed to be enediyne clusters if each of the five genes was present and within 40 open reading frames (ORFs) of each other. The abundance of available *Salinispora* genomes in the JGI database threatened a bias towards *Salinispora* enediynes during GNN generation and analysis. We therefore chose seven *Salinispora* genomes (six of the seven strains contained two enediyne gene clusters each) as

representatives and removed the other 180 clusters representing 120 *Salinispora* genomes. Analysis of the enediyne gene clusters from these six *Salinispora* genomes found they were reasonable representatives for the *Salinispora* enediyne gene clusters. After dereplication of the *Salinispora* genomes, a total of 87 potential enediyne gene clusters from 78 different bacteria strains were identified, collected, and used to construct the enediyne GNN.

With 87 enediyne gene clusters publicly available, yet only 15 structurally characterized enediyne natural products, 10 of which have sequenced gene clusters, it is evident that enediynes are underexplored and therefore provide great optimism that the field of enediyne chemistry is still in its infancy. Several phyla of bacteria are equipped with the necessary genetic information to produce enediynes with actinobacteria as the most prolific producers with 68 members out of the 78 representative strains (Table S1). This is consistent with the fact that all but two (namenamicin and shishijimicin) of the structurally characterized enediynes were isolated from actinobacteria. Given the complexities associated with enediyne biosynthesis, an effective strategy to analyze, predict, and prioritize potential enediyne gene clusters for production and biosynthetic studies is needed to ensure the greatest chance of success of discovering new enediyne natural products and elucidating their biosyntheses.

Enediyne GNNs as a model for GNN functionality for natural product analysis and prediction

A list of proteins translated from the genes within the genome neighborhoods of the *pksE* cassettes of each potential enediyne producer, specifically ±40 ORFs from the *pksE*, were collected. Thus, each genome neighborhood consisted of 81 proteins. Due to the incomplete sequencing of a few genomes, enediyne gene clusters on small scaffolds gave genome neighborhoods consisting of <81 proteins. We considered that putative gene clusters of 81 ORFs would be more than adequate to assess the general nature of conserved and diverse proteins found in enediyne biosynthesis. The *pksE* was chosen as the central gene as it is the first committed step for enediyne polyketide biosynthesis [18,55]. Choosing ±40 ORFs from the *pksE* gave an unbiased view of each genome neighborhood. It should be noted, however, that most of the known enediyne gene clusters contain their *pksE*s near the boundaries of their gene cluster (Fig. 2A). This may result in irrelevant proteins included in the GNN, but their unrelatedness will likely separate them from genes encoding enediyne biosynthetic proteins, confirming them as background noise. Proteins <70 amino acids in length, which are unlikely to have biosynthetic functions, were removed from each genome neighborhood. Given their short lengths and consequently large E-values, these proteins would have likely appeared as singletons in the GNN. The total combined list of proteins (5814) was used in an all vs. all BLAST using BLAST+. Both self-loops (i.e., A vs A) and duplicates (i.e., A vs B and B vs A) were deleted from the BLAST result. A preliminary E-value threshold of 1.0×10^{-8} was established based on the lowest found similarity, and therefore largest E-value, between CagA, the apoprotein of C-1027 consisting of 143 amino acids, and any putative apoproteins in the population. Even at this low stringency threshold, almost every class of proteins found in the enediyne GNN are separated (Fig. 3A). The few exceptions include, but are not limited to, the E2/E3 family and the PKSE with other non-enediyne PKSs. A

smaller E-value threshold, and thus higher stringency, would allow the separation of these related families.

The ability of GNNs [57], and similarity networks in general [4], to overlay orthogonal information corresponding to each protein used in the analysis, creates additional opportunities to find and display trends in a visual and thought-provoking manner. Pertinent orthogonal information for natural product biosynthesis gene clusters may include members of known gene clusters, predicted or experimentally determined protein functions, subgroups of natural product families (e.g., 9- vs 10-membered enediynes), taxonomy, and genetic proximity (e.g., gene distance from *pksE*). We constructed three different enediyne GNNs at an E-value threshold of 10^{-8} to highlight the utility of this feature.

The enediyne GNN (Fig. 3A) highlights the structurally characterized enediynes with sequenced gene clusters. Each node (protein) from the known enediyne members is depicted based on its classification as 9- (large circle) or 10-membered (large diamond) cores with unique colors based on the enediyne it helps to produce. Each node is additionally labeled with its corresponding name or ORF number, and functionally characterized proteins (Table S2), by either in vitro or in vivo experiments, are outlined (bold, red lines). It is quickly evident which protein families are conserved for all enediynes, conserved based on associated peripheral moieties, unique to certain enediyne structures, and novel or unrelated to enediyne biosynthesis. As shown with protein similarity networks [4,8,56], outlining the nodes corresponding to functionally characterized proteins can help identify and prioritize protein families for future studies. For example, several genes (i.e., *sgcJ* and *sgcM* homologues) are fairly well conserved throughout both known and putative enediyne gene clusters, but their functions remain unknown.

The same enediyne GNN can be modified to represent strain information or genetic location. Figure 4 depicts the most representative genera from actinobacteria as well as the other phyla. While most protein families contain members from several different genera, there are also genus-specific protein families; the NcsR2/NcsR3 family from *Streptomyces* and several *Amycolatopsis* families are examples of genus-specific proteins. Figure 5 was constructed to highlight the proteins with the closest genetic proximity to the *pksE* gene. One would expect if a gene is in close proximity to its *pksE*, there is a greater likelihood that the gene in question is also involved in enediyne biosynthesis. A set of conserved “core” enediyne proteins (E2, E3, E4, E5, E6, E7, and E10) whose encoding genes are typically within five ORFs of the *pksE* is evident (Fig. 5). Other genes, while still highly conserved, show greater variation in their gene locations. The networks shown in this study are just a few examples of how natural products gene clusters can be analyzed and visualized using GNNs; nevertheless, the effectiveness of GNNs to probe natural product biosynthesis is clear.

GNNs identifying enediyne-associated chemistry

During the analysis of these GNNs (Figs. 3–5), we were able to quickly assign proteins as enediyne-essential or enediyne-associated while simultaneously eliminating non-enediyne “noise.” A set of highly conserved enediyne proteins present in the majority of the 87 gene clusters appears to be essential for the biosynthesis of both 9- and 10-membered enediynes.

Only the proteins encoded by the enediyne cassette, *E3/E4/E5/pksE/E10*, were present in all 87 gene clusters with the partially sequenced gene cluster for ESP being the one exception. It is, therefore, tempting to predict that the essential genes for enediyne core biosynthesis are contained exclusively within the enediyne cassette. Other enediyne-associated proteins showing high levels of conservation include homologues of the C-1027 proteins (Sgc) B, D2, E2 (clustered with E3), E6, E7, E8, E9, E11, F, J, L, M, and R2, as well as the calicheamicin proteins (Cal) R2 and T5. The functions, or putative functions (Table S2), of these proteins are likely involved in the divergence of 9- and 10-membered cores or for regulation and resistance. While a subset of these protein families (B, E2/E3, E7, R2, CalR2, and CalT5) contained known members of both 9- and 10-membered enediynes (Fig. 3B), several protein families do not contain CAL or DYN homologues (D2, E6, E8, E9, E11, F, J, L, M) and thus appear to be 9-membered specific (Fig. 3C). In contrast, there are only a few protein families (CalR3, CalT4/DynT8, S6, and U20) that do not contain known 9-membered homologues, appearing to be 10-membered specific (Fig. 3D). The less conserved protein families and nonconserved proteins are likely unrelated to enediynes altogether. While there are examples of singletons or doubletons with relevance to specific enediynes, prioritization to study the conserved protein families will impact enediyne chemistry as a whole.

Each enediyne core is decorated with peripheral moieties (Fig. 1); examples include aminoglycosides (C-1027, NCS, MDP, KED, CAL, ESP), anthraquinones (DYN, UCM), β-amino acids (C-1027, MDP, KED), benzoxazolinates (C-1027), naphthoates (NCS, KED), and phenolic acid derivatives (MDP, CAL, ESP). A closer look at the key proteins involved in the biosynthesis of these moieties in the GNN predicts the peripheral moieties in each putative enediyne (Table S1). For example, *Streptomyces* sp. CNS615 possesses key homologues for the ligations of benzoxazolinate (SgcD6), β-amino acid (SgcC5), and naphthoate (NcsB4) moieties to its enediyne core (Fig. S1). An enediyne with these decorations would be novel as no known enediyne contains each of these three moieties. Analysis of all the genes in the genome neighborhood of the *pksE* from *S.* sp. CNS615 reveals it possesses all the necessary genes for benzoxazolinate biosynthesis, but is missing the amino lyase (SgcC4) and halogenase (SgcC3) from the β-amino acid of C-1027 and the naphthoate PKS (NcsB). This result suggests that two of the peripheral moieties on this enediyne are different from known moieties, the proteins are sufficiently different to fall into different protein families, or a GNN of ±40 ORFs is insufficient for complete analysis of the enediyne gene cluster from *Streptomyces* sp. CNS615.

GNNs simplifying the prediction of 9- and 10-membered enediynes

The inability to use bioinformatics to accurately predict if a gene cluster produces 9- or 10-membered enediynes has slowed the discovery of novel enediyne natural products. A major reason is the current lack of known 10-membered enediyne gene clusters as only the CAL and DYN gene clusters have been reported. A small sample set of two, both from *Micromonospora* spp., makes it extremely difficult to identify trends associated with 10-membered enediyne chemistry. For example, CalU16 and CalU19 were recently reported as self-sacrifice resistance proteins for the CALs [12]. No homologous proteins are found in the DYN cluster suggesting that the CalU16/U19 family of resistance proteins is not

conserved in all 10-membered enediynes and therefore cannot be used as a 10-membered indicator (Fig. 3E).

With almost 300 putative enediyne gene clusters in the public databases, we anticipated a larger sample set (87 used in this study) would allow the identification of indicators for both 9- and 10-membered enediynes. Understandably, the apoproteins associated with only 9-membered enediynes seemed to be an appropriate indicator, yet only nine of the 87 gene clusters contained homologous proteins at an E-value of 10^{-8} (Fig. 3E). At this E-value, it was unclear which proteins are directly responsible for the differentiation between the 9- and 10-membered enediyne scaffolds. As discussed above, the C-1027 homologues D2, E6, E8, E9, E11, F, J, L, and M appear to be 9-membered specific. Without knowing the functions of these proteins, with the exception of the flavin reductase SgcE6 [51] and the epoxide hydrolase SgcF [19,28], we attempted to use these genes as indicators for 9-membered enediynes in the 10^{-8} GNN. By assigning putative gene clusters as 9- or 10-membered based on each indicator, we anticipated some protein families to be grouped together by enediyne core size. A GNN depicting a resultant prediction based on color quickly displays if the indicator is of high quality. The use of a good indicator (protein family) of 9-membered enediynes as a prediction criterion would leave a good indicator (protein family) of 10-membered enediynes untouched. Using the 10^{-8} GNN resulted in no high quality indicators as a 9-membered prediction based on the proteins listed above failed to reveal conserved protein families that are 10-membered specific.

One advantage of GNNs [57] and protein similarity networks [4], is the ability to adjust the E-value threshold to find highly related protein families. Increasing the stringency of the threshold (i.e., smaller E-value) begins to pull apart distantly related protein families and results in more, smaller families. At an E-value of 10^{-75}, we noticed the E2/E3 cluster pulled apart to reveal E2 and E3 as discrete subfamilies (Fig. 6B). While the E3 subfamily is a member of the enediyne cassette and contains known members of both 9- and 10-membered enediynes, the E2 subfamily did not contain homologues from the known 10-membered enediynes (i.e., CAL and DYN), and thus appeared to be 9-membered specific. Using the new E2 subfamily as a 9-membered indicator left two conserved protein families untouched. The two families, protein kinases, including CalR3, DynR3, and DynORF18, and ABC-type transporters, including CalT6 and CalT7, are presumed regulatory and resistance proteins, respectively (Table S2). At an E-value of 10^{-75}, the CalR3 family consisted of more members (27) than the CalT6/T7 family (17); therefore, a GNN using the CalR3 family as a 10-membered indicator was constructed (Fig. 7). This simple prediction strategy, the E2 family for 9-membered and the CalR3 family for 10-membered, resulted in 47 (seven known) and 25 (three known) putative 9- and 10-membered enediynes, respectively; fourteen gene clusters, without E2 or CalR3 homologues, were left unpredicted (Table S1). Of the 14 unpredicted gene clusters, only five are from actinobacteria. Several of the unpredicted gene clusters from actinobacteria possess homologues in other protein families that contain only 9-membered enediynes implying these are also 9-membered enediynes and reinforcing the utility of GNNs for natural product prediction (Table S1). Of the non-actinobacteria enediyne producers, nine of 11 could not be predicted using this strategy, suggesting these indicators are most indicative of the enediynes from actinobacteria. Interestingly, a subcluster of the E3 family in the GNNs at an E-value of

10^{-75} contains 10 of the 11 nonactinobacteria (Fig. 7B, the E3 protein from *Bacillus* is a singleton at this threshold), suggesting nonactinobacteria contain conserved protein families that may be used as alternative indicators. Phylogenetic analysis of the E2/E3 family of proteins supported that E2s are distinct from three groups of E3s (Fig. 8). The three groups include "typical" E3s comprised of a large clade and two smaller clades of proteins with close relationships to the E2 clade, and two divergent clades of "atypical" E3s from actinobacteria and nonactinobacteria. The atypical clade of actinobacteria all contain homologues of CalR3, suggesting this clade is exclusively involved in 10-membered enediyne biosynthesis. DynU15 is phylogenetically distinct from all other E2 and E3s.

CONCLUSION

With the ability to utilize GNNs to predict 9- and 10-membered enediynes as well as the peripheral moieties decorating the enediyne cores, strain and gene cluster prioritization for novel enediyne discovery, or rediscovery of known enediynes with improved characteristics, is now possible in a high-throughput manner. Typically, once a gene cluster of interest is found, the gene-encoded proteins are individually annotated based on their homologous proteins in the database, which may or may not be associated with a related natural product. The collected functions of the entire gene cluster are then compared with those of known natural products. GNNs facilitate gene cluster annotation using all related natural product gene clusters in a single step. The high-throughput nature of GNNs will become even more useful as the number of sequenced genomes increases. The construction of an enediyne GNN and the examples discussed in this study should inspire further investigations into the enediyne family of natural products. While the enediynes are an excellent example of the utility of GNNs for natural products discovery and biosynthesis, it is clear that this bioinformatics tool is applicable to any natural product gene cluster of interest.

MATERIALS AND METHODS

Virtual survey of enediyne gene clusters

Using phmmer (http://hmmer.janelia.org/search/phmmer) and the C-1027 PKSE (SgcE) as the query, we searched for homologous PKSE proteins in the nonredundant (NR) database. Protein hits with lengths between 1700 and 2200 amino acids with E-values $<10^{-50}$ were collected. The proteins comprising the rest of the enediyne PKS cassette in C-1027, SgcE3, SgcE4, SgcE5, and SgcE10, were also individually used as search queries and protein hits with E-values $<10^{-50}$ were also collected. If one strain contained each of the five proteins and the distances between each of the PKS cassette-encoding genes were <40 ORFs, it was counted as a potential enediyne producer. For the genome sequences deposited in the Integrated Microbial Genomes (IMG) system of the JGI database (http://img.jgi.doe.gov/#IMGSystems), the same five proteins (SgcE, SgcE3, SgcE4, SgcE5, and SgcE10) were used as queries to run blastp against all the proteins in the bacteria domain with the sequencing status "all finished, permanent draft and draft." Potential enediyne producers from the JGI database were determined in the same manner as for the NR database. The combined list of strains and gene clusters from the NR and JGI databases were combined and dereplicated to create an initial unique list of potential enediyne producers. Due to an

abundance of *Salinispora* genomes in the JGI database, two representatives of each of the three *Salinispora* species, *arenicola*, *pacifica*, and *tropica*, were chosen; the remaining *Salinispora* gene clusters were deleted.

Genome neighborhood network (GNN)

Genes within the genome neighborhoods of the *pksE* cassettes of each potential enediyne producer, specifically ±40 ORFs from the *pksE*, were collected and translated. Thus, each genome neighborhood consisted of 81 proteins, unless incomplete sequencing prevented a full genome neighborhood construction. Proteins <70 amino acids in length, which are unlikely to have biosynthetic functions, were removed from the protein collection. The total combined list of proteins (5814) was used in an all vs. all BLAST using BLAST+ [6], the BLOSUM62 matrix [49], and an E-value limit of 10. Both self-loops (i.e., A vs A) and duplicates (i.e., A vs B and B vs A) were deleted from the BLAST result. Before GNN generation, the E-values for each comparison (input sequences) were converted into integers using –log(E-value) for easier threshold management in Cytoscape. E-values of 0 were arbitrarily converted into 200, i.e., a larger value than the largest calculated value. Cytoscape v3.0 was used for GNN generation, visualization, and analysis [44]. All GNNs were displayed using the “organic” layout. Access to the enediyne GNN is available at www.scripps.edu/shen/NPLI/database.html.

Phylogenetic analysis of the E2 and E3 families

The E2 and E3 sequences were aligned with ClustalW using Bioedit 7.2.5 [16]. A maximum likelihood phylogenetic tree was generated using the Jones-Taylor-Thornton (JTT) model of amino acid substitution and 1000 bootstrap replications in MEGA 6.06 [48].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work is supported in part by National Institutes of Health Grant CA78747 and the Natural Products Library Initiative at the Scripps Research Institute.

REFERENCES

1. Ahlert J, Shepard E, Lomovskaya N, Zazopoulos E, Staffa A, Bachmann BO, Huang K, Fonstein L, Czisny A, Whitwam RE, Farnet CM, Thorson JS. The calicheamicin gene cluster and its iterative type I enediyne PKS. Science. 2002; 297:1173–1176. [PubMed: 12183629]
2. Belecki K, Crawford JM, Townsend CA. Production of octaketide polyenes by the calicheamicin polyketide synthase CalE8: implications for the biosynthesis of enediyne core structures. J Am Chem Soc. 2009; 131:12564–12566. [PubMed: 19689130]
3. Boghaert ER, Sridharan L, Armellino DC, Khandke KM, DiJoseph JF, Kunz A, Dougher MM, Jiang F, Kalyandrug LB, Hamann PR, Frost P, Damle NK. Antibody-targeted chemotherapy with the calicheamicin conjugate hu3S193-*N*-acetyl γ calicheamicin dimethyl hydrazide targets Lewisy and eliminates Lewisy-positive human carcinoma cells and xenografts. Clin Cancer Res. 2004; 10:4538–4549. [PubMed: 15240546]
4. Brown SD, Babbitt PC. Inference of functional properties from large-scale analysis of enzyme superfamilies. J Biol Chem. 2012; 287:35–42. [PubMed: 22069325]

5. Buchanan GO, Williams PG, Feling RH, Kauffman CA, Jensen PR, Fenical W. Sporolides A and B: structurally unprecedented halogenated macrolides from the marine actinomycete *Salinispora tropica*. Org Lett. 2005; 7:2731–2734. [PubMed: 15957933]
6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden Thomas L. BLAST+: architecture and applications. BMC Bioinf. 2009; 10:421.
7. Chari RVJ. Targeted Cancer Therapy: conferring specificity to cytotoxic drugs. Acc Chem Res. 2008; 41:98–107. [PubMed: 17705444]
8. Chow J-Y, Tian B-X, Ramamoorthy G, Hillerich BS, Seidel RD, Almo SC, Jacobson MP, Poulter CD. Computational-guided discovery and characterization of a sesquiterpene synthase from *Streptomyces clavuligerus*. Proc Natl Acad Sci USA. 2015; 112:5661–5666. [PubMed: 25901324]
9. Davies J, Wang H, Taylor T, Warabi K, Huang X-H, Andersen RJ. Uncialamycin, a new enediyne antibiotic. Org Lett. 2005; 7:5233–5236. [PubMed: 16268546]
10. Dedon PC, Goldberg IH. Sequence-specific double-strand breakage of DNA by neocarzinostatin involves different chemical mechanisms within a staggered cleavage site. J Biol Chem. 1990; 265:14713–14716. [PubMed: 2144279]
11. Edo K, Mizugaki M, Koide Y, Seto H, Furihata K, Otake N, Ishida N. The structure of neocarzinostatin chromophore possessing a novel bicyclo[7.3.0]dodecadiyne system. Tet Lett. 1985; 26:331–334.
12. Elshahawi SI, Ramelot TA, Seetharaman J, Chen J, Singh S, Yang Y, Pederson K, Kharel MK, Xiao R, Lew S, Yennamalli RM, Miller MD, Wang F, Tong L, Montelione GT, Kennedy MA, Bingman CA, Zhu H, Phillips GN, Thorson JS. Structure-guided functional characterization of enediyne self-sacrifice resistance proteins, CalU16 and CalU19. ACS Chem Biol. 2014; 9:2347–2358. [PubMed: 25079510]
13. Galm U, Hager MH, Van Lanen SG, Ju J, Thorson JS, Shen B. Antitumor antibiotics: bleomycin, enediynes, and mitomycin. Chem Rev. 2005; 105:739–758. [PubMed: 15700963]
14. Gao Q, Thorson JS. The biosynthetic genes encoding for the production of the dynemicin enediyne core in *Micromonospora chersina* ATCC53710. FEMS Microbiol Lett. 2008; 282:105–114. [PubMed: 18328078]
15. Golik J, Dubay G, Groenewold G, Kawaguchi H, Konishi M, Krishnan B, Ohkuma H, Saitoh K, Doyle TW. Esperamicins, a novel class of potent antitumor antibiotics. 3. Structures of esperamicins A1, A2, and A1b. J Am Chem Soc. 1987; 109:3462–3464.
16. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999; 41:95–98.
17. Hirama M, Akiyama K, Tanaka T, Noda T, Iida K-i, Sato I, Hanaishi R, Fukuda-Ishisaka S, Ishiguro M, Otani T, Leet JE. Paramagnetic enediyne antibiotic C-1027: spin identification and characterization of radical species. J Am Chem Soc. 2000; 122:720–721.
18. Horsman GP, Chen Y, Thorson JS, Shen B. Polyketide synthase chemistry does not direct biosynthetic divergence between 9- and 10-membered enediynes. Proc Natl Acad Sci USA. 2010; 107:11331–11335. [PubMed: 20534556]
19. Horsman GP, Lechner A, Ohnishi Y, Moore BS, Shen B. Predictive model for epoxide hydrolase-generated stereochemistry in the biosynthesis of nine-membered enediyne antitumor antibiotics. Biochemistry. 2013; 52:5217–5224. [PubMed: 23844627]
20. Jones RR, Bergman RG. p-Benzyne. Generation as an intermediate in a thermal isomerization reaction and trapping evidence for the 1,4-benzenediyl structure. J Am Chem Soc. 1972; 94:660–661.
21. Kennedy DR, Ju J, Shen B, Beerman TA. Designer enediynes generate DNA breaks, interstrand cross-links, or both, with concomitant changes in the regulation of DNA damage responses. Proc Natl Acad Sci USA. 2007; 104:17632–17637. [PubMed: 17978180]
22. Kobayashi S, Ashizawa S, Takahashi Y, Sugiura Y, Nagaoka M, Lear MJ, Hirama M. The First Total Synthesis of N1999-A2: Absolute Stereochemistry and Stereochemical Implications into DNA Cleavage. J Am Chem Soc. 2001; 123:11294–11295. [PubMed: 11697974]
23. Komano K, Shimamura S, Norizuki Y, Zhao D, Kabuto C, Sato I, Hirama M. Total synthesis and structure revision of the (-)-maduropeptin chromophore. J Am Chem Soc. 2009; 131:12072–12073. [PubMed: 19655742]

24. Kong R, Goh LP, Liew CW, Ho QS, Murugan E, Li B, Tang K, Liang Z-X. Characterization of a carbonyl-conjugated polyene precursor in 10-membered enediyne biosynthesis. J Am Chem Soc. 2008; 130:8142–8143. [PubMed: 18529057]
25. Lane AL, Nam S-J, Fukuda T, Yamanaka K, Kauffman CA, Jensen PR, Fenical W, Moore BS. Structures and comparative characterization of biosynthetic gene clusters for cyanosporasides, enediyne-derived natural products from marine Actinomycetes. J Am Chem Soc. 2013; 135:4171–4174. [PubMed: 23458364]
26. Lee MD, Dunne TS, Chang CC, Ellestad GA, Siegel MM, Morton GO, McGahren WJ, Borders DB. Calichemicins, a novel family of antitumor antibiotics. 2. Chemistry and structure of calichemicin γ1I. J Am Chem Soc. 1987; 109:3466–3468.
27. Liang Z-X. Complexity and simplicity in the biosynthesis of enediyne natural products. Nat Prod Rep. 2010; 27:499–528. [PubMed: 20336235]
28. Lin S, Horsman GP, Chen Y, Li W, Shen B. Characterization of the SgcF epoxide hydrolase supporting an (*R*)-vicinal diol intermediate for enediyne antitumor antibiotic C-1027 biosynthesis. J Am Chem Soc. 2009; 131:16410–16417. [PubMed: 19856960]
29. Liu W, Ahlert J, Gao Q, Wendt-Pienkowski E, Shen B, Thorson JS. Rapid PCR amplification of minimal enediyne polyketide synthase cassettes leads to a predictive familial classification model. Proc Natl Acad Sci USA. 2003; 100:11959–11963. [PubMed: 14528002]
30. Liu W, Christenson SD, Standage S, Shen B. Biosynthesis of the enediyne antitumor antibiotic C-1027. Science. 2002; 297:1170–1173. [PubMed: 12183628]
31. Liu W, Nonaka K, Nie L, Zhang J, Christenson SD, Bae J, Van Lanen SG, Zazopoulos E, Farnet CM, Yang CF, Shen B. The neocarzinostatin biosynthetic gene cluster from *Streptomyces carzinostaticus* ATCC 15944 involving two iterative type I polyketide synthases. Chem Biol. 2005; 12:293–302. [PubMed: 15797213]
32. Lohman JR, Huang S-X, Horsman GP, Dilfer PE, Huang T, Chen Y, Wendt-Pienkowski E, Shen B. Cloning and sequencing of the kedarcidin biosynthetic gene cluster from *Streptoalloteichus* sp. ATCC 53650 revealing new insights into biosynthesis of the enediyne family of antitumor antibiotics. Mol BioSyst. 2013; 9:478–491. [PubMed: 23360970]
33. McDonald LA, Capson TL, Krishnamurthy G, Ding W-D, Ellestad GA, Bernan VS, Maiese WM, Lassota P, Discafani C, et al. Namenamicin, a new enediyne antitumor antibiotic from the marine ascidian *Polysyncraton lithostrotum*. J Am Chem Soc. 1996; 118:10898–10899.
34. McGlinchey RP, Nett M, Moore BS. Unraveling the biosynthesis of the sporolide cyclohexenone building block. J Am Chem Soc. 2008; 130:2406–2407. [PubMed: 18232689]
35. Myers AG. Proposed structure of the neocarzinostatin chromophore-methyl thioglycolate adduct; a mechanism for the nucleophilic activation of neocarzinostatin. Tet Lett. 1987; 28:4493–4496.
36. Myers AG, Fraley ME, Tom NJ, Cohen SB, Madar DJ. Synthesis of (+)-dynemicin A and analogs of wide structural variability: establishment of the absolute configuration of natural dynemicin A. Chem Biol. 1995; 2:33–43. [PubMed: 9383401]
37. Nam S-J, Gaudencio SP, Kauffman CA, Jensen PR, Kondratyuk TP, Marler LE, Pezzuto JM, Fenical W. Fijiolides A and B, inhibitors of TNF-α-induced NFκB activation, from a marine-derived sediment bacterium of the genus *Nocardiopsis*. J Nat Prod. 2010; 73:1080–1086. [PubMed: 20481500]
38. Oh D-C, Williams PG, Kauffman CA, Jensen PR, Fenical W. Cyanosporasides A and B, chloro- and cyano-cyclopenta[a]indene glycosides from the marine actinomycete "*Salinispora pacifica*". Org Lett. 2006; 8:1021–1024. [PubMed: 16524258]
39. Oku N, Matsunaga S, Fusetani N. Shishijimicins A-C, novel enediyne antitumor antibiotics from the ascidian *Didemnum proliferum*. J Am Chem Soc. 2003; 125:2044–2045. [PubMed: 12590521]
40. Otani T, Yoshida K-I, Sasaki T, Minami Y. C-1027 enediyne chromophore: presence of another active form and its chemical structure. J Antibiot. 1999; 52:415–421. [PubMed: 10395278]
41. Ren F, Hogan PC, Anderson AJ, Myers AG. Kedarcidin chromophore: synthesis of its proposed structure and evidence for a stereochemical revision. J Am Chem Soc. 2007; 129:5381–5383. [PubMed: 17417855]

42. Rudolf JD, Dong L-B, Huang T, Shen B. A genetically amenable platensimycin- and platencin-overproducer as a platform for biosynthetic explorations: a showcase of PtmO4, a long-chain acyl-CoA dehydrogenase. Mol BioSyst. 2015
43. Senter PD. Potent antibody drug conjugates for cancer therapy. Curr Opin Chem Biol. 2009; 13:235–244. [PubMed: 19414278]
44. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]
45. Shen B, Hindra, Yan X, Huang T, Ge H, Yang D, Teng Q, Rudolf JD, Lohman JR. Enediynes: exploration of microbial genomics to discover new anticancer drug leads. Bioorg Med Chem Lett. 2015; 25:9–15. [PubMed: 25434000]
46. Sievers EL, Appelbaum FR, Spielberger RT, Forman SJ, Flowers D, Smith FO, Shannon-Dorcy K, Berger MS, Bernstein ID. Selective ablation of acute myeloid leukemia using antibody-targeted chemotherapy: a phase I study of an anti-CD33 calicheamicin immunoconjugate. Blood. 1999; 93:3678–3684. [PubMed: 10339474]
47. Sugimoto Y, Otani T, Oie S, Wierzba K, Yamada Y. Mechanism of action of a new macromolecular antitumor antibiotic, C-1027. J Antibiot. 1990; 43:417–421. [PubMed: 2161819]
48. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013; 30:2725–2729. [PubMed: 24132122]
49. Thompson JD, Higgins DG, Gibson TJ. Improved sensitivity of profile searches through the use of sequence weights and gap excision. CABIOS, Comput Appl Biosci. 1994; 10:19–29. [PubMed: 8193951]
50. Thorson JS, Sievers EL, Ahlert J, Shepard E, Whitwam RE, Onwueme KC, Ruppen M. Understanding and exploiting nature's chemical arsenal: the past, present and future of calicheamicin research. Curr Pharm Des. 2000; 6:1841–1879. [PubMed: 11102565]
51. Van Lanen SG, Lin S, Horsman GP, Shen B. Characterization of SgcE6, the flavin reductase component supporting FAD-dependent halogenation and hydroxylation in the biosynthesis of the enediyne antitumor antibiotic C-1027. FEMS Microbiol Lett. 2009; 300:237–241. [PubMed: 19817865]
52. Van Lanen SG, Oh T-j, Liu W, Wendt-Pienkowski E, Shen B. Characterization of the maduropeptin biosynthetic gene cluster from *Actinomadura madurae* ATCC 39144 supporting a unifying paradigm for enediyne biosynthesis. J Am Chem Soc. 2007; 129:13082–13094. [PubMed: 17918933]
53. Van Lanen SG, Shen B. Biosynthesis of enediyne antitumor antibiotics. Curr Topics Med Chem. 2008; 8:448–459.
54. Zazopoulos E, Huang K, Staffa A, Liu W, Bachmann BO, Nonaka K, Ahlert J, Thorson JS, Shen B, Farnet CM. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nature Biotechnol. 2003; 21:187–190. [PubMed: 12536216]
55. Zhang J, Van Lanen SG, Ju J, Liu W, Dorrestein PC, Li W, Kelleher NL, Shen B. A phosphopantetheinylating polyketide synthase producing a linear polyene to initiate enediyne antitumor antibiotic biosynthesis. Proc Natl Acad Sci USA. 2008; 105:1460–1465. [PubMed: 18223152]
56. Zhang X, Kumar R, Vetting MW, Zhao S, Jacobson MP, Almo SC, Gerlt JA. A unique cis-3-hydroxy-L-proline dehydratase in the enolase superfamily. J Am Chem Soc. 2015; 137:1388–1391. [PubMed: 25608448]
57. Zhao S, Jacobson Matthew P, Sakai A, Zhang X, Kumar R, San Francisco B, Solbiati J, Gerlt John A, Vetting Matthew W, Hillerich B, Seidel Ronald D, Almo Steven C, Steves A, Brown S, Akiva E, Barber A, Babbitt Patricia C. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. eLife. 2014; 3:e03275.
58. Zhen Y, Ming X, Yu B, Otani T, Saito H, Yamada Y. A new macromolecular antitumor antibiotic, C-1027. III. Antitumor activity. J Antibiot. 1989; 42:1294–1298. [PubMed: 2759910]

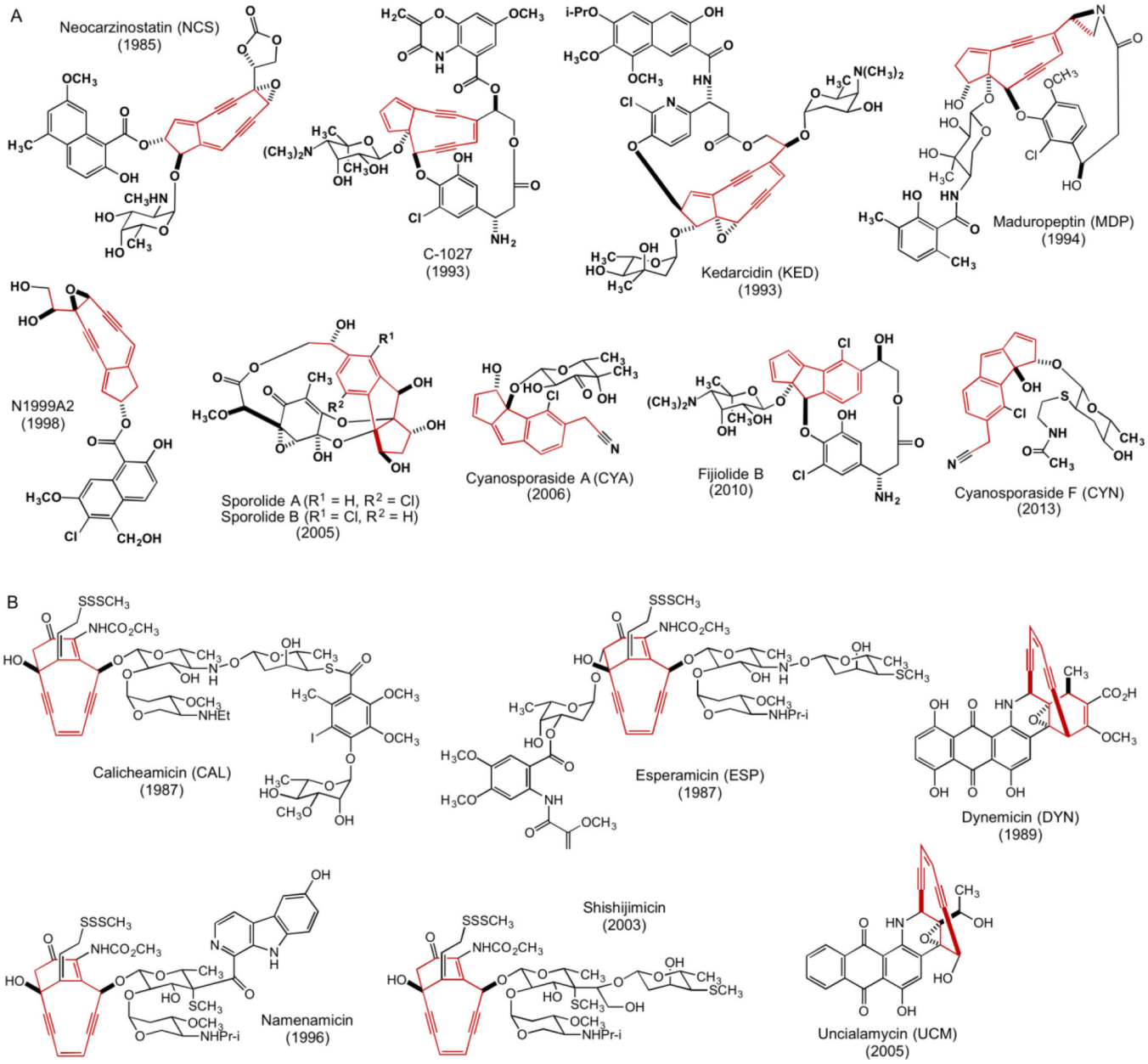

Fig. 1.

Structures of the 11 known enediynes, (A) five 9-membered and (B) six 10-membered representatives, with their enediyne cores highlighted in red. The sporolides, cyanosporasides, and fijolides were proposed to be derived from 9-membered enediynes. Given in parentheses are the years when each enediyne structure was established.

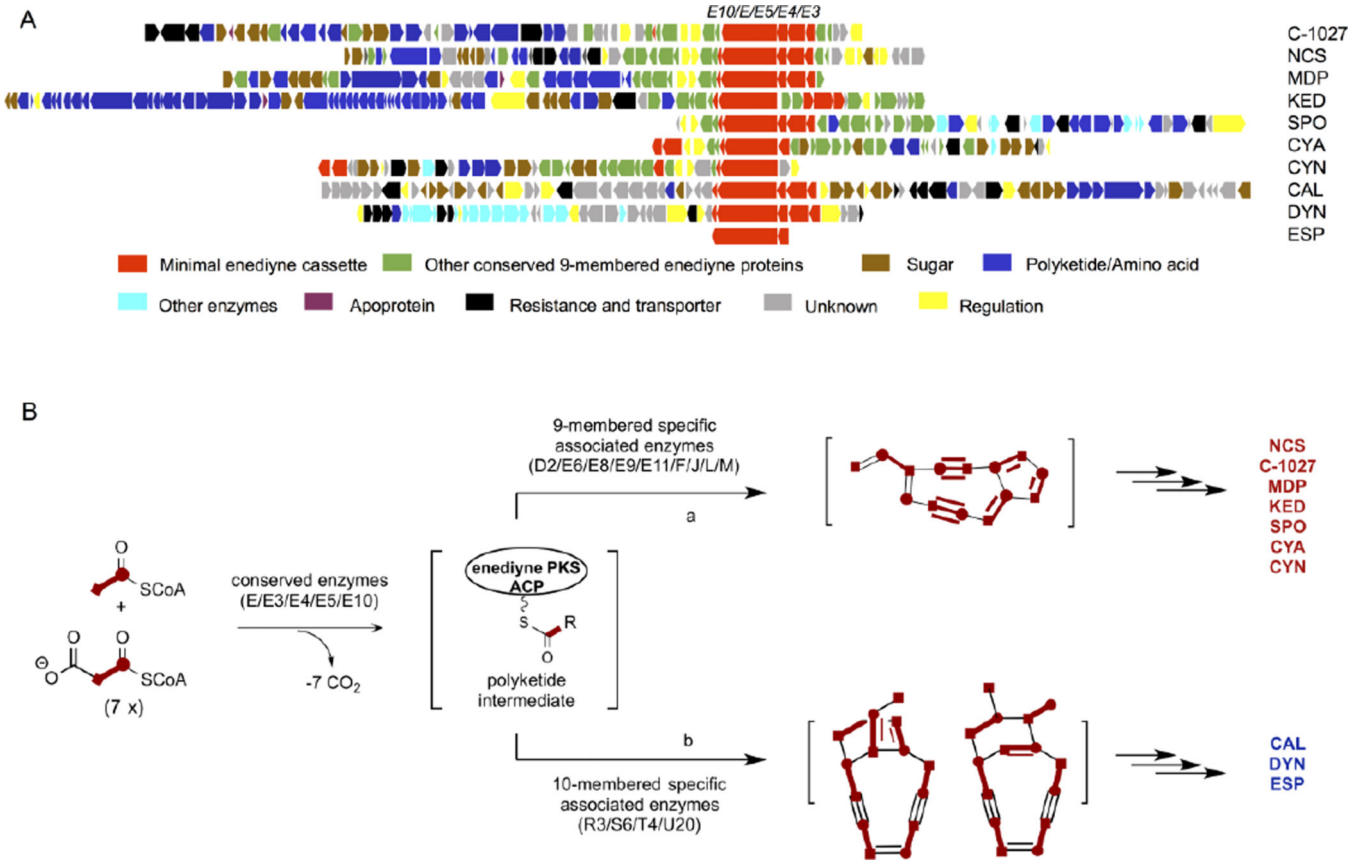

Fig. 2.

Characterization of the enediyne biosynthetic machinery providing an opportunity to explore bacterial genomes for the discovery of new enediyne natural products. (A) Ten known enediyne clusters highlighting the *pksE* cassettes (i.e., *E3*, *E4*, *E5*, *E*, *E10*, red genes) common to all enediynes and used as a beacon for enediyne producers. The surrounding ORFs in the gene clusters, forming the *pksE* genome neighborhoods, are color-coded to signify their involvement in peripheral moiety biosynthesis, pathway regulation, and self-resistance. (B) Unified model for enediyne core biosynthesis. Conserved proteins in all enediynes (E, E3, E4, E5, and E10) produces an ACP-tethered or free polyketide intermediate that is converted into the proposed enediyne cores in the presence of 9- (path a) or 10-membered (path b) associated enzymes. Examples of 9- and 10-membered associated enzymes are named according to the homologues of C-1027 and CAL, respectively.

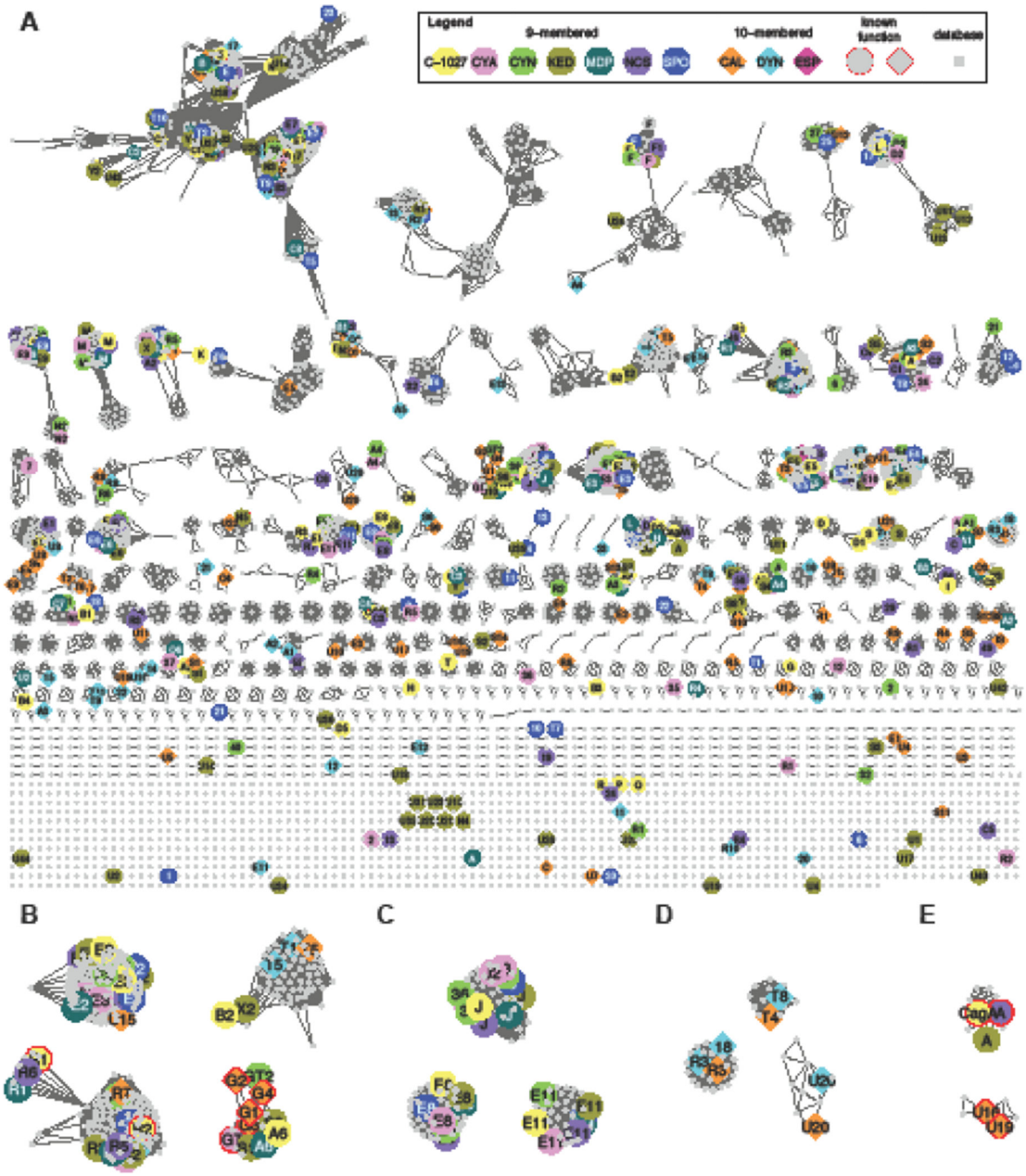

Fig. 3.

Genome neighborhood network (GNN) for the enediyne family of natural products. (A) The GNN displayed with an E-value threshold of 10^{-8}. Each node is colored and shaped based on the enediyne it produces, labeled with its corresponding gene name or ORF number, and highlighted if it has been functionally characterized (see inset legend). (B–D) Selected families of conserved proteins involved in both 9- and 10-membered, only 9-membered, or only 10-membered enediyne biosynthesis, respectively. (E) Members of the sequestration apoprotein family for 9-membered enediynes and the self-sacrifice protein family for CAL.

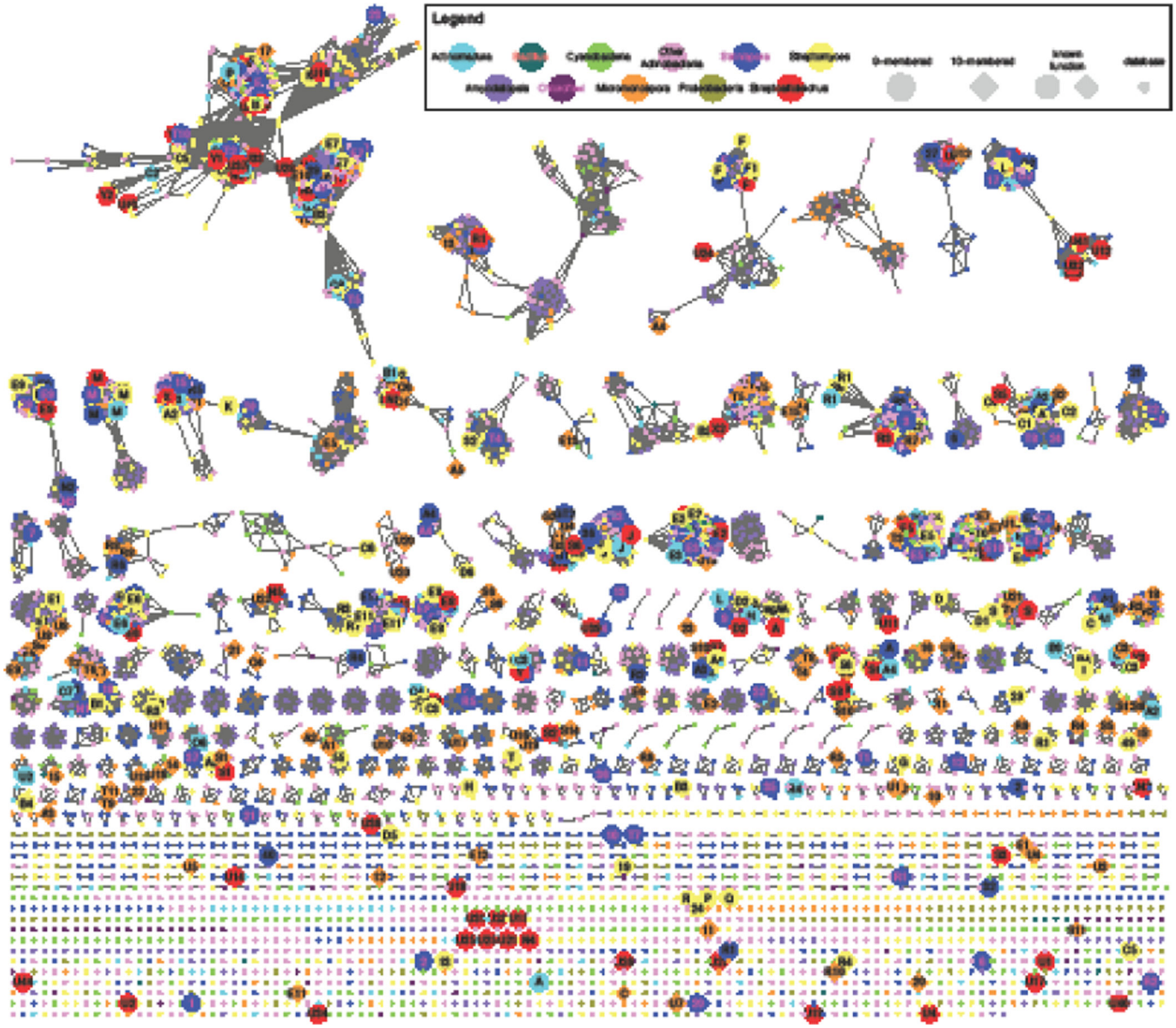

Fig. 4.

GNN depicting genera of enediyne producers. The GNN displayed with an E-value threshold of 10^{-8}. Each node is colored based on the taxonomic identification of the strain it is found in, shaped based on the size of the enediyne core it produces, labeled with its corresponding gene name or ORF number, and highlighted if it has been functionally characterized (see inset legend).

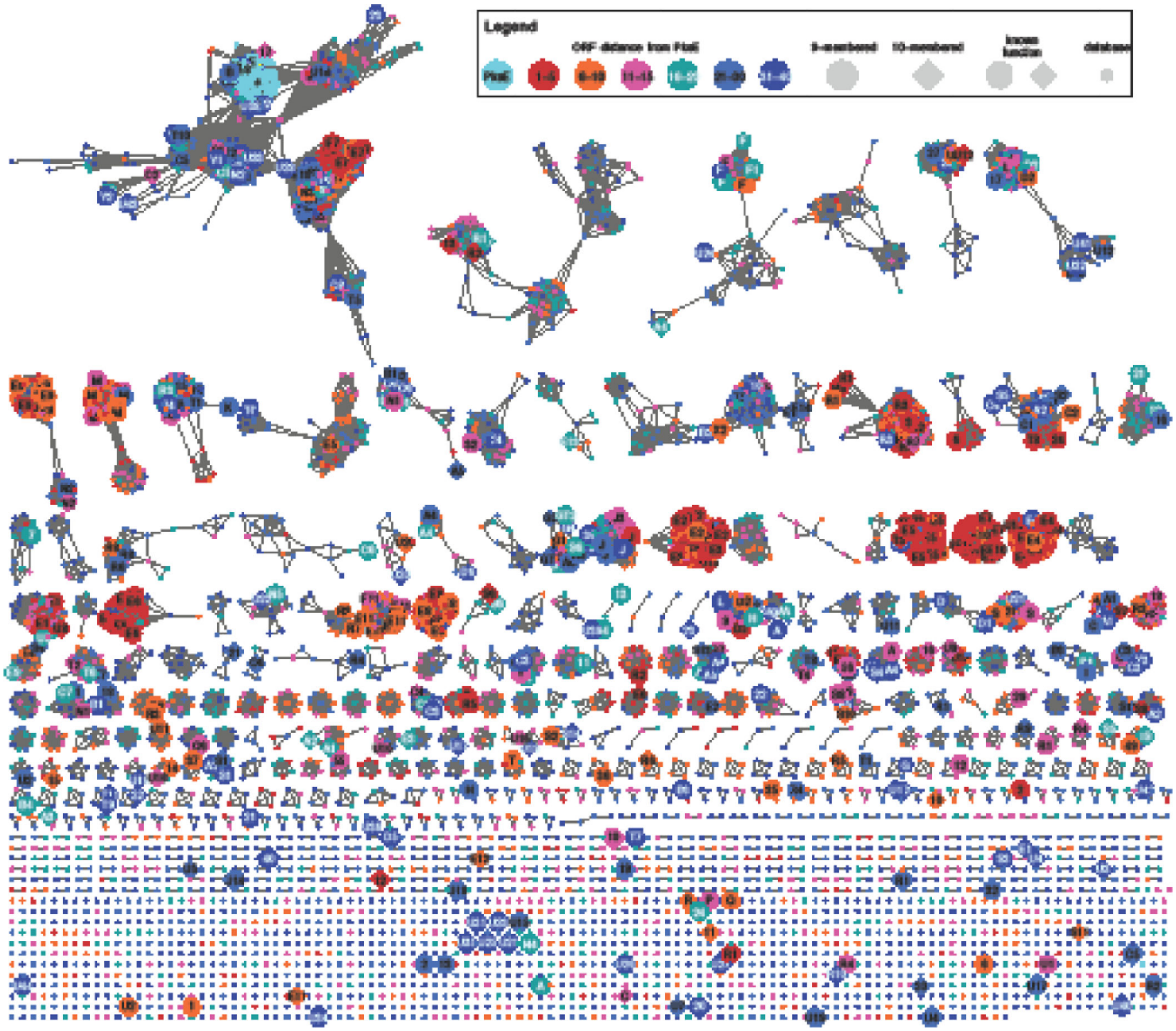

Fig. 5.

GNN depicting genetic location in reference to *pksE*. The GNN displayed with an E-value threshold of 10^{-8}. Each node is colored based on its genetic distance from its *pksE*, shaped based on the size of the enediyne core it produces, labeled with its corresponding gene name or ORF number, and highlighted if it has been functionally characterized (see inset legend).

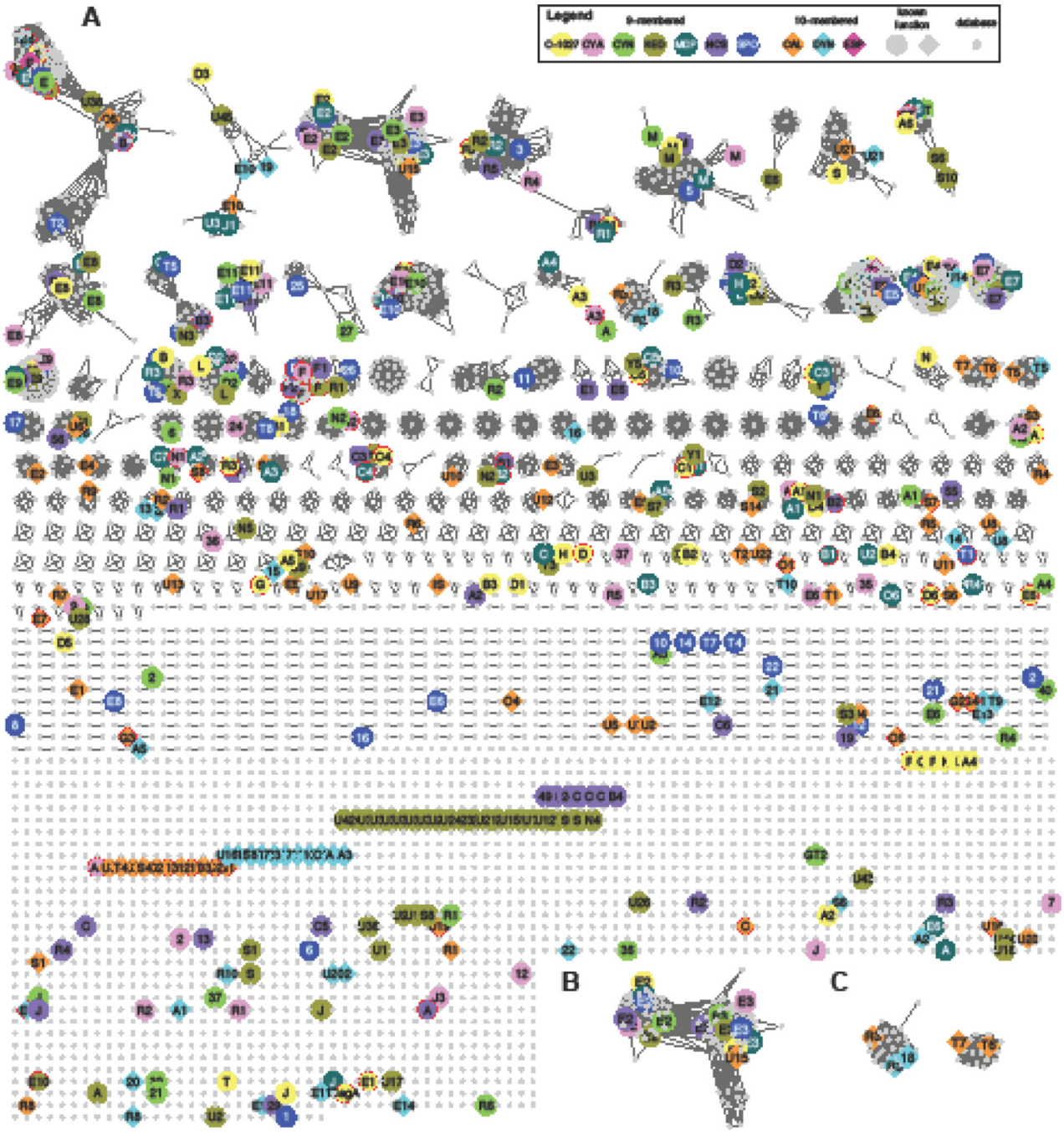

Fig. 6.

High stringency GNN for the enediyne family of natural products. (A) The GNN displayed with an E-value threshold of 10^{-75}. Each node is colored and shaped based on the enediyne it produces, labeled with its corresponding gene name or ORF number, and highlighted if it has been functionally characterized (see inset legend). For comparisons of GNNs at a lower threshold or with 9- or 10-membered enediyne prediction, see Figs. 3 or 7, respectively. (B) The E2 and E3 family is separated into three groups (see Fig. 3B for comparison at an E-

value threshold of 10^{-8}). (C) Families of conserved proteins (i.e., CalR3 and CalT6/T7) from gene clusters not containing an E2 homologue.

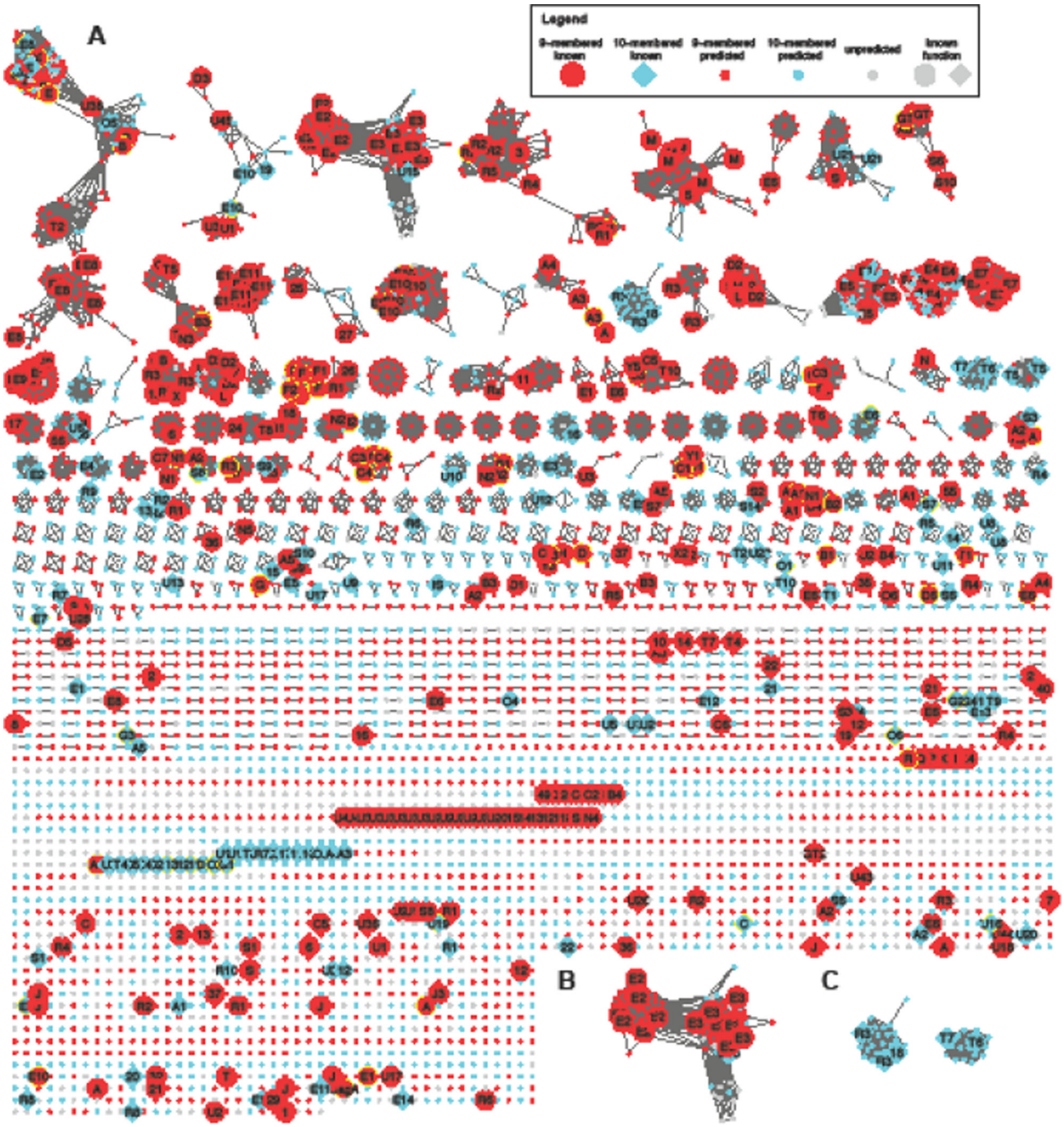

Fig. 7.

Genome neighborhood network (GNN) facilitates the prediction of 9- and 10-membered enediynes. (A) The GNN displayed with an E-value threshold of 10^{-75}. Using the E2 and CalR3 families for 9- and 10-membered enediyne indicators, respectively, each node is colored based and shaped on the size of the enediyne core it produces or is predicted to produce (see inset legend). Each node is labeled with its corresponding gene name or ORF number, and highlighted if it has been functionally characterized. See Fig. 6 for a complementary GNN colored based on the known enediyne it produces. See Fig. S2 for a

GNN displaying the predicted 9- and 10-membered enediynes at an E-value threshold of 10^{-8}. (B) The E2 and (C) CalR3 families as 9- and 10-membered enediyne indicators, respectively.

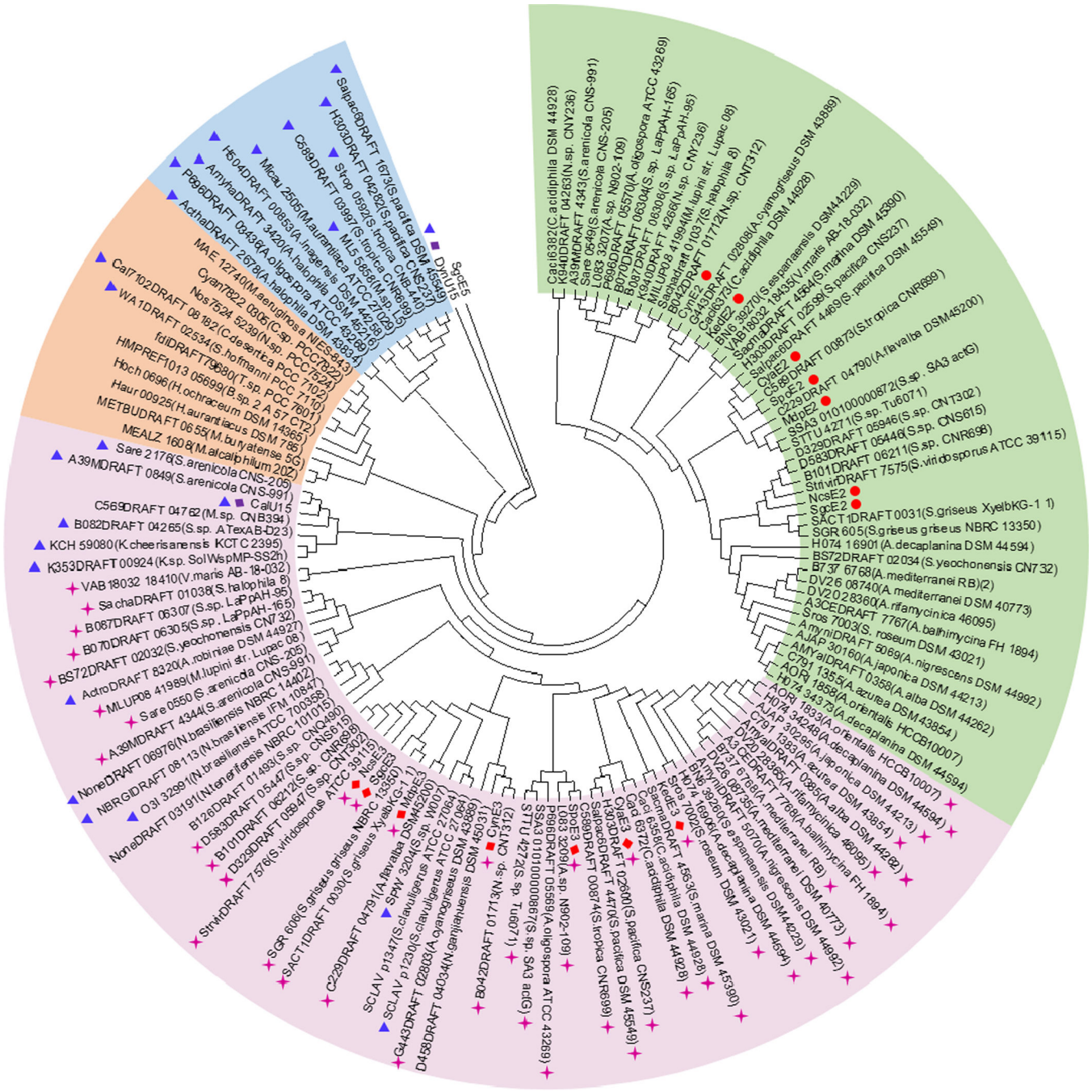

Fig. 8.

Phylogenetic analysis of the E2 and E3 family of proteins conserved in enediyne gene clusters. The E2 and E3 proteins are separated into four groups: E2s (green), typical E3s with close relationships to the E2 family (pink), atypical E3s from nonactinobacteria (orange), and atypical E3s from actinobacteria (blue). DynU15 is phylogenetically distinct. Each protein is labeled with its locus tag with its corresponding bacteria strain in parentheses. The E2 and E3 proteins from known 9-membered enediyne gene clusters are represented with red dots and red diamonds, respectively. The E3 proteins from known 10-

membered enediyne gene clusters are represented with purple diamonds. E3 proteins with E2 or CalR3 homologues in their gene clusters are highlighted with red stars or blue triangles, respectively. SgcE5 was used as an outgroup.