NetSyn: prokaryotic genomic context exploration of protein families

Mark Stam[1], Jordan Langlois[1], Céline Chevalier[1], Jean Mainguy[1], Guillaume Reboul[1], Karine Bastard[1], Claudine Médigue[1], David Vallenet[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, Evry, France

Corresponding Author: Mark Stam, mstam@genoscope.cns.fr

Abstract

Background: The growing availability of large prokaryotic genomic datasets presents an opportunity to discover new metabolic pathways and enzymatic reactions useful for industrial or synthetic biological applications. Efforts to identify new enzyme functions in this vast number of sequences cannot be achieved without bioinformatics tools and the development of new strategies. Standard methods for assigning a biological function to a gene are based on sequence similarity. However, complementary approaches rely on mine databases to identify conserved gene clusters (i.e. syntenies). In prokaryotic genomes, genes involved in the same pathway are frequently encoded in a single locus with an operonic organisation. This genomic context conservation is considered as a reliable indicator of functional relationships, and is therefore a promising approach for improving ~~the~~ gene function prediction.

Methods. Here we present NetSyn (Network Synteny), a tool to group protein sequences based on the conservation of their genomic context rather than solely on sequence similarity. From a list of protein sequence identifiers, NetSyn searches corresponding genome entries to retrieve neighboring genes. Corresponding protein sequences are grouped into families to define homology relationships and compute a synteny conservation score between the different extracted genomic contexts. A network is then created in which the nodes represent the input proteins and the edges indicate that two proteins share a conserved synteny. Finally, the network is partitioned into clusters grouping proteins with similar genomic contexts, using a community detection algorithm.

Results. As a proof of concept, we used NetSyn on two different datasets. The first one is the BKACE protein family (formerly named DUF849) which has previously been divided into isofunctional sub-families. NetSyn was able to go a step further by providing additional sub-families beyond those already described. The second dataset corresponds to a set of non-homologous proteins belonging to three different glycoside hydrolase (GH) families. These GHs are known to work cooperatively in a Polysaccharide-Utilization Loci (PUL) and are

therefore grouped together in the same genomic contexts. NetSyn was able to identify a locus grouping 3 GHs, involved in the degradation of xyloglucan, in 162 prokaryotic genomes.

Discussion. By highlighting conserved synteny in distantly related prokaryotic species, NetSyn enables functional links between proteins to be established beyond sequence similarity alone. We showed that NetSyn is efficient for exploring large prokaryotic protein families, enabling the definition of isofunctional groups and the identification of functional interactions between non-homologous enzymes. These features enable the prediction of new genomic structures that have not yet been experimentally characterized.

Finally, NetSyn is also useful for pinpointing annotation errors that have been propagated across databases, and for suggesting annotations on proteins lacking functional prediction. NetSyn is freely available at https://github.com/labgem/netsyn.

Introduction

Advances in genome sequencing technologies have created an ever-increasing gap between the availability of protein sequences and the slow progress in their functional characterization and ~~thus,~~ annotation (Galperin and Koonin, 2010). It is currently estimated that 23% of the families defined in the PFAM database are annotated as “unknown function” (Mistry et al., 2021).

This issue is further compounded by automatic annotation procedures, which still exhibit a high false-positive rate. In particular for methods that rely solely on sequence similarity, leading to annotation error rates of up to 80% for certain protein families (Goudey et al., 2022; Rembeza and Engqvist, 2021; Schnoes et al., 2009; Steinegger and Salzberg, 2020). Conversely, many known enzymatic activities have not yet been associated with protein sequences. In their review, Sorokina *et al.* show that more than 1000 EC numbers, referred to as orphan reactions, are not linked to a specific gene (Sorokina et al., 2014). To overcome the limitations of these methods, innovative strategies have been developed such as a combination of genome analysis and metabolic information (Chen et al., 2011; Kharchenko et al., 2006; Smith et al., 2012; Yamanishi et al., 2007), an integrative approach associating sequence family classification and controlled vocabulary (Jung et al., 2014), a Protein Protein Interaction (PPI) network (Cozzetto et al., 2013; Peng et al., 2014; Zhao et al., 2016), and more recently, a machine learning technique based on natural language processing methods (Ofer et al., 2021).

One way to address the challenge of annotating gene function is to compare gene genomic contexts which is considered a reliable indicator of functional relationships (Huynen and Snel, 2000; Janga et al., 2005; McClean et al., 2010; Rogozin et al., 2002) and has been used to support functional annotation (Ferrer et al., 2010; Lee et al., 2016; Vallenet et al., 2006). Genomic context methods can be classified into four categories (Ferrer et al., 2010): gene cluster, phylogenetic profile, gene fusion (or Rosetta Stone) and gene neighbour. The gene cluster

method (or conserved synteny) is defined as the conservation of chromosomal proximity between genes (Overbeek et al., 1999; Tamames, 2001). Based on conserved genes in several organisms, the cluster method computes distances, in number of bases, between two adjacent genes transcribed in the same direction (Bowers et al., 2004; Overbeek et al., 1999). Small distances are considered to be a good indicator of co-regulated gene groups, as observed in operon structure (Brouwer et al., 2008) or Polysaccharide Utilisation Loci (PUL) (Bjursell et al., 2006). The phylogenetic profile method (Pellegrini et al., 1999) is based on the comparison of ~~the~~ proteomes across genomes: genes A and B are functionally related if orthologs of A and B show similar patterns of presence and absence across many genomes. The gene fusion method (Marcotte et al., 1999) is based on the gene fusion events that can occur in some genomes. Pairs of proteins linked by such a fusion event are more likely to be functionally related and involved in the same biological process, for example, within the same metabolic pathway. The gene neighbour method (Bowers et al., 2004) calculates the distance between two genes in different genomes, based on the number of genes that appear between the two homologs plus one. Genes A and B are functionally related if orthologs of A and B across many organisms are located in close proximity on the chromosome (also called gene neighbours). Careful selection of genomes is required to avoid taxonomic bias (*e.g.,* organisms from the same genus).

Further efforts have been made to incorporate gene context information into integrative approaches. For instance, the STRING database (Szklarczyk et al., 2017) stores known and predicted protein-protein interactions from different resources. These interactions can be physical (*i.e.* heterodimer) or indirect (*e.g.* gene regulation, signal relay mechanism) and are represented as a graph in which nodes correspond to proteins and edges to interactions. Such networks capture interactions between different protein families. However, the analyses are limited by the genomes included in the database and cannot support the exploration of an entire protein family. Other tools have been developed to search for conserved genomic contexts. For instance, Cblaster (Gilchrist et al., 2020) and CAGECAT (van den Belt et al., 2023), use the protein sequences of target genes and their genomic context as input data to perform a BLAST search in the NCBI reference genome database. Only blast hits separated by a maximum distance defined by the user (in bp) and belonging to the same organism and scaffold are retained. Another example is given by GCsnap (Pereira, 2021), which retrieves the genomic contexts by searching for the flanking genes of the input sequences; then it compares the retrieved genomic context. The Enzyme Function Initiative (EFI) (Oberg et al., 2023; Zallot et al., 2019) has also led to the development of EFI-GNT, a method based on the Genomic Neighborhood Network (GNN), which enables the analysis of up to hundreds of genomes and generate large-scale datasets. It computes a Sequence Similarity Network (SSN) of the input protein sequences, in which groups of highly connected nodes are considered as a cluster. Then, neighbouring genes of each protein are retrieved from the European Nucleotide Archive (ENA) (Yuan et al., 2024) and considered related if they belong to the same PFAM family (Mistry et al., 2021). These methods do not rely

on the creation of a true genomic context network, compared to a recent method called Syntenet (Almeida-Silva et al., 2023; Tao, 2025; Zhao and Schranz, 2017). Starting from whole-genome protein sequences, it searches for blocks of up to 50 genes conserved between two genomes and calculates a score for each of these blocks. A network is then created where the nodes represent genes in syntenic blocks and the edges represent syntenic relationships between genes. This network is divided using community detection algorithms and can be used to highlight synteny relationships. Syntenet is able to infer and analyze micro-synteny networks. It has been successfully applied on eukaryotic genomes (Guo et al., 2025; Li et al., 2025; Zhao and Schranz, 2019).

A strategy based partly on the search for conserved genomic contexts has enabled the division of the DUF849 family into isofunctional subgroups (Bastard et al., 2014). These results led us to develop NetSyn (Network Synteny), a tool able to detect conserved syntenies within distantly related organisms. Designed to analyse families from prokaryotic genomes, it links proteins into a network based on the conservation of their genomic context, highlighting specific loci involved in the same biological process. NetSyn provides several features to explore the synteny network, such as sequence clustering, visualisation of the genomic context of each input sequence, and integration of data from various resources. Overall, our tool quantifies the conservation of the genomic context of proteins from different prokaryotic genomes.

Compared to the tools mentioned above, NetSyn focuses on the genomic context of members of the analysed protein families to calculate a synteny score between their genomic context. It has been validated on two types of protein datasets. The first one is the initial set of sequences used for the analysis of the β-keto acid cleavage enzyme (BKACE) family (Bastard et al., 2014). The second one consists of sequences homologous to the three glycoside hydrolases present in a PUL that degrades xyloglucan. Overall, the NetSyn is based on the principle that proteins sharing a similar genomic context across distantly related bacterial genomes are more likely to share similar functions or be involved in the same biological function (e.g. a metabolic pathway). Consequently, our method is well suited for partitioning diverse prokaryotic protein families into isofunctional clusters.

Materials & Methods

NetSyn workflow

NetSyn is divided into four steps (Fig.1): i) Genomic context extraction, ii) Protein family computation, iii) Synteny computation and scoring, iv) Synteny network computation and clustering (Figure S1 illustrates how the computation time of NetSyn scales with the number of input sequences).

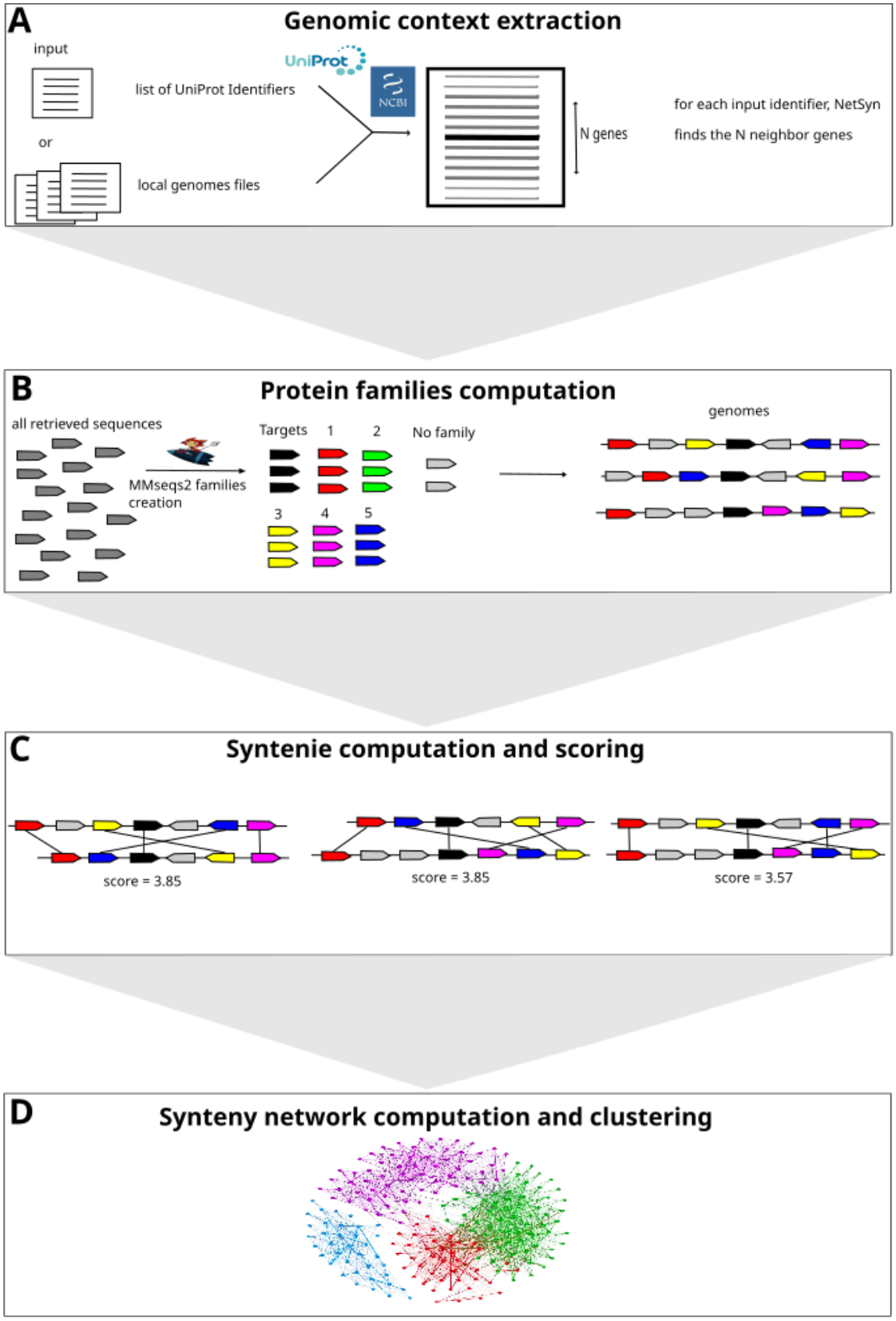

A
Genomic context extraction
input
list of UniProt Identifiers
or
local genomes files
UniProt
NCBI
N genes
for each input identifier, NetSyn
finds the N neighbor genes
B
Protein families computation
all retrieved sequences
MMseqs2 families
creation
Targets
1
2
3
4
5
No family
genomes
C
Syntenie computation and scoring
score = 3.85
score = 3.85
score = 3.57
D
Synteny network computation and clustering

Figure 1 : *NetSyn Workflow*

From a list of UniProt accessions, NetSyn downloads the corresponding genome assembly files or the user can provide their own assembly files. NetSyn then parses these files to recover the protein sequences of the target genes and their neighboring genes (panel A). The retrieved protein sequences are clustered to define protein families (default values : 30% of identity on 80% of coverage) (panel B). Synteny relationships and scores are then computed for each pair of target proteins (panel C). A network is generated in which target sequences are represented as nodes, and edges are created when a synteny with a score above a threshold (3 by default) is found. Finally, the network is partitioned using four different community detection algorithms (panel D).

Genomic context extraction

NetSyn takes as input a list of UniProt accession numbers (hereafter referred to as target proteins). The workflow uses the UniProt rest API to retrieve, for each entry, the corresponding nucleotide accession number and to download the associated genome assembly file (hereafter called the genome file) from the NCBI. Alternatively, NetSyn can accept local genome files listed in a tab-delimited file containing the following mandatory fields:

- **protein_AC**: protein target accession number in the genome file
- **protein_AC_field**: target protein accession number in the genome file
- **nucleic_AC**: nucleotide sequence accession number of the genome file
- **nucleic_File_Format**: format of the genome file (EMBL or Genbank)
- **nucleic_File_Path**: path to the locally stored genome file

Moreover, a metadata file containing attributes of the target proteins (*e.g.* a PFAM or INTERPRO family classification, the taxonomy of the corresponding organism) can be provided. These attributes are displayed on the final synteny network (see below).

From the genome files, the sequence of the target proteins and their genomic context are extracted. The number of neighbouring genes to include can be adjusted according to the desired context size; by default, this value is set to 11 (i.e. 5 genes upstream and 5 genes downstream of the target gene). As the average number of genes in prokaryotic operons range from 2 to 3, this window size is generally sufficient to detect most conserved gene clusters (Zheng et al., 2002).

NetSyn also retrieves the taxonomic lineage of the organisms from which the input sequences originate by querying the NCBI API using the taxonomic identifier extracted from the genome file. As a result, the taxonomic lineage provided by NetSyn is derived from the NCBI and is regularly curated (Schoch et al., 2020). This taxonomic information can be mapped onto the final network and can also be used to remove redundancy (see section “Merging nodes and network reduction”).

Protein family computation

The purpose of this step is to define homology relationships between proteins. All the retrieved proteins (target proteins and proteins from neighboring genes) are clustered using MMseqs2 (Steinegger and Söding, 2017). A common rule of thumb is that proteins sharing at least 30% sequence identity over at least 80% of their length are likely derived from a common ancestor (Rost, 1999). We used these values as default values (these parameters can be modified). The resulting clusters enable NetSyn to define a set of homologous protein families which are found in the context of the target proteins. These families are then used to compute synteny conservation for each pair of target proteins (Fig. 1).

Synteny computation and scoring

Genomic context conservation (*i.e*. synteny conservation) between two target proteins is computed using an exact graph-theoretical approach (Boyer et al., 2005). From the two genomes containing the target proteins (denoted T_A and T_B), two networks (N_A and N_B) are constructed containing the target proteins and their neighboring genes as vertices ($G_{A1\text{-}n}$ and $G_{B1\text{-}n}$, according to the window size parameter), with edges representing their colocalization along the genome. To account for potential gaps (*i.e.* genes within a conserved synteny that lack homologs in the other genomes), a partial transitive closure is applied to N_A and N_B, linking genes separated by no more than n genes (synteny gap parameter, set to 3 by default). Edges are then added between N_A and N_B to represent homology relationships between genes belonging to the same protein families computed previously. Finally, a common connected component (CCC) between the two networks is identified by constructing a multigraph in which vertices from N_A and N_B are merged based on homology relationships. These CCCs correspond to conserved synteny groups. Figure 2 illustrates the method, showing the extracted genomic contexts around T_A and T_B (panel A) and the resulting detected synteny (panel B). G_{A1} and G_{B2} genes are not part of the conserved synteny because the number of genes between G_{A1} and T_A in the genome A exceed the gap parameter threshold.

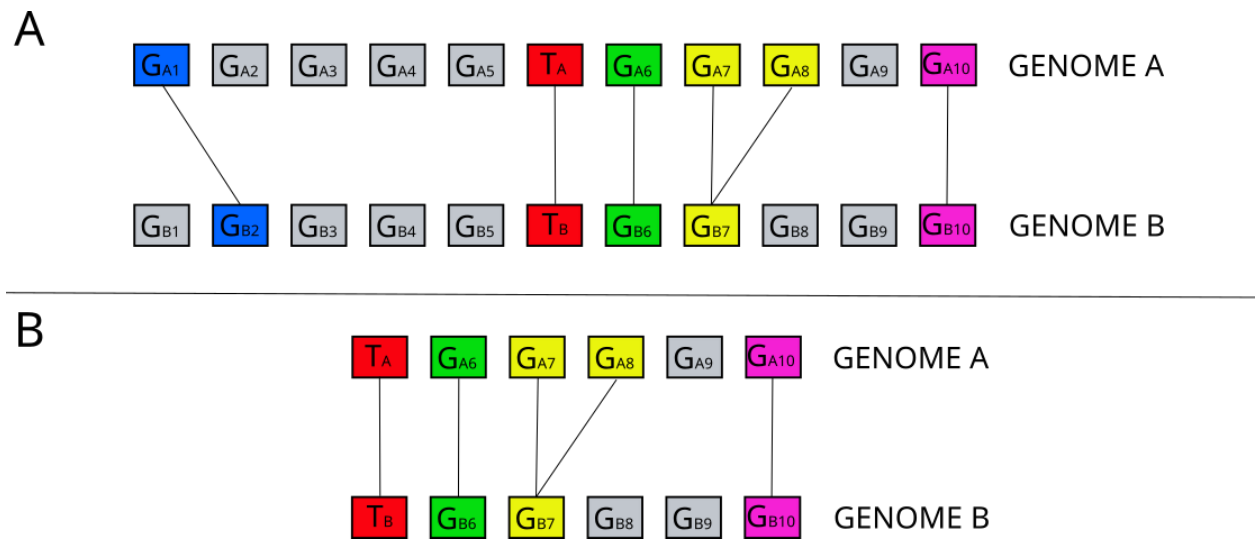


Figure 2: *Example of synteny computation between two target genes.*
A) Genomic context extraction from two target genes: T_A from genome A and T_B from genome B. In this example the window size is set to 11, meaning that NetSyn considers 5 genes upstream and 5 genes downstream of each target gene to extract the genomic contexts in both genomes. Homologous genes between the two genomic contexts are represented by coloured rectangles and linked by edges.
B) Detected synteny. The synteny predicted by NetSyn includes the following pairs of homologous genes: T_A/T_B, G_{A6}/G_{B6}, G_{A7}/G_{B7}, G_{A8}/G_{B7}, G_{A10}/G_{B10}. The G_{A1}/G_{B2} homologs are not part of the conserved synteny because the number of genes between G_{A1} and T_A in genome A (i.e. 4 genes) is greater than the gap parameter threshold (set to 3 by default). 9 genes are conserved in the synteny out of 12 genes, including 3 gap genes (G_{A9}, G_{B8} and G_{B9}). The Synteny Score is thus equal to (9/2)(9/12)=3.375.*

Once a conserved synteny has been identified for a pair of target proteins, a score is calculated as follow:

$$Synteny\ Score = (GS/2) * (GS/GT)$$

where GS is the number of genes involved in the synteny group, and GT is the total number of genes in both genomic contexts, including gap genes. This score represents the average number of genes in the synteny, penalized by the presence of gap genes. For example, in Figure 2 (panel B), the genomic context contains 9 genes in synteny among out of a total of 12 genes including 3 gap genes (G_{A9}, G_{B8} and G_{B9}). Therefore the Synteny Score (SS) is equal to (9/2) * (9/12) = 3.375.

Synteny network computation and clustering

Once all syntenies and scores have been computed between target proteins, NetSyn builds a synteny network in which nodes represent the target proteins. Two target proteins are connected by an edge if a conserved synteny is detected between them. The weight of edges are equal to synteny scores, highlighting the most conserved syntenies. To retain only significant syntenies, NetSyn, applies a default threshold of 3 on the synteny score, although this value can be adjusted

by the user. Target proteins that are not connected to any other target proteins are excluded from the final network.

To group target proteins sharing similar genomic contexts, several community detection algorithms are implemented in the NetSyn workflow: MCL, Walktrap, Louvain and Infomap. The Markov Cluster algorithm (Dongen, 2000) (MCL) is an unsupervised algorithm based on the simulation of flow in graphs (it simulates random walks within a graph). The Walktrap algorithm (Pons and Latapy, 2005) is also based on random walks and performs well on large networks (> 1000 vertices). The Louvain algorithm (Blondel et al., 2008) is based on the comparison of the edge density inside and outside a community (i.e. a cluster). The Infomap algorithm (Rosvall and Bergstrom, 2008) tries to minimise the length of a randomised walk trajectory into a cluster.

These four algorithms are independently applied to identify groups of target proteins sharing similar genomic contexts. In the synteny network, the resulting clusters are visualized using distinct colors, with each color representing a group of related target proteins. By default, the clustering method producing the smallest number of clusters is displayed in the HTML output. Our analyses indicate that this method frequently corresponded to the one that minimized the formation of 'small' and non pertinent clusters (i.e., clusters containing between 2 and 5 nodes). However, users can also explore the results obtained with other clustering methods, both in the HTML output and in network visualisation software.

Cluster quality

In the clustering framework based genomic context conservation, a cluster is considered coherent when most nodes within the cluster are interconnected. To assess this property, we use the alpha index which gives a measure of connectivity within a given cluster. The alpha index is the ratio between the number of cycles (i.e., closed paths that start and end at the same node and passing through nodes only once) and the estimated total number of cycles in the network. It is calculated using the following formula: alpha = (e-v)/((v(v-1)/2)-(v-1)) where e is the number of edges and v is the number of vertices. A value of 1 indicates a fully connected cluster while a value of 0 means a completely disconnected cluster (i.e., a minimally connected cluster without cycles).

NetSyn Output

We provide two output files for exploring and analysing the network : a graphML file which can be opened with network visualisation tools such as Gephi or Cytoscape, and an HTML file that can be viewed in a Web browser (Fig. 3). In the graphML output, information from the metadata file and the extracted taxonomic lineage are mapped onto the network.

Figure 3: Visualisation of the NetSyn results.

Panel A displays the NetSyn network. The control panel allows customization of node colors based on the "Attributes of targets"selection, which includes NetSyn clusters (obtained with one of the four available algorithms), taxonomic ranks, or user-defined metadata. Nodes can be selected using the "Search by" functionality, with the following search fields: Labels (target genes names), Products (functional descriptions), ECs (annotated EC numbers), NetSyn family IDs (protein family identifiers). In this panel, the network can be zoomed in or out and rotated.
Panel B provides the color legend for the currently selected 'Attributes of targets' ordered by decreasing number of protein targets. The user can click on a specific coloured item (e.g., cluster 1 obtained with the walktrap method) to highlight the corresponding nodes in the network.
Panel C displays the genomic context of the selected cluster. In this view, each arrow represents a gene and homologous genes (i.e genes belonging to the same protein family) share the same color. Target genes are marked with a golden star. When the user hovers over a gene, a pop-up window displays the following information: protein family identifier, strain name, UniProt accession, locus tag, product, gene name, and EC number (if available).
Panel D displays a summary table of the protein families identified within the selected cluster. The table includes the following columns: NetSyn family ID (protein family identifier), Synteny Nb (the number of occurrences of the family in synteny), Proteins in Synteny Nb (number of corresponding proteins from the family in syntenies), Species Nb (the number of species), Strain Nb (the number of strains), Genome Nb (the number of genomes), Product (the different annotations for the proteins of the family) ordered by frequency of occurrences.

The HTML view is divided into four panels. The synteny network is displayed in the central panel (A), where each node represents a target protein, and an edge between two nodes indicates that a conserved genomic context with a synteny score above a given threshold (3 by default) has been identified between them. Nodes can be colored according to different attributes, such as cluster identifiers, genome taxonomy or user-defined metadata. The control panel at the top of

the A panel allows users to select the attributes and information displayed on the synteny network (see Fig. 3 captions). The left panel (B) displays the network legend, which depends on the parameters selected in the control panel, the identifier, the number of protein targets and the alpha index (see below) of each cluster (Fig. 3B). Selecting a coloured item in panel B, automatically highlights the corresponding nodes in the synteny network (panel A). In addition, the genomic contexts of the corresponding target proteins are displayed on the right panel (C). Each context is centered on the target protein gene, which is marked with a yellow star, surrounded by neighboring genes belonging to the conserved synteny. In this view, homologous genes share the same colour (*i.e.*, they belong to the same protein family). The bottom panel (D) provides, for the selected group of target proteins (panel B), a summary table of the protein families found within the conserved syntenies, including the number of proteins, the number of corresponding species, strains and genomes, as well as functional annotations.

Two additional files are also created:

- A <output directory name>_Report_1.txt file, which summarizes the results of the different steps (e.g., number of identifiers without corresponding Genbank ID or assembly file, number of clusters obtained with each clustering method, etc.).
- A <output dir name>_Settings_1.yaml file, which records the parameters used to run NetSyn.

Merging nodes and network reduction

In some cases, the final network can be large and difficult to analyse. To address this issue, we implemented an option to reduce the network. When using this option, a single network clustering method and a specific property must be selected (i.e. a column name in the metadata file, or a taxonomic rank retrieved by NetSyn). NetSyn then performs a standard run, but nodes belonging to the same NetSyn cluster and sharing the given property are merged into a single node. As NetSyn can be influenced by the taxonomic relatedness, we recommend using the taxonomic rank genus to merge nodes (see SI 2). Therefore, the “merged nodes” method is the result of the intersection between NetSyn clusters (obtained with the selected clustering method) and one property. The resulting reduced network retains the overall structure of the original network while containing fewer nodes.

Results & Discussion

The NetSyn method was evaluated on two datasets. The first dataset consists of a family of homologous proteins, the β-keto acid cleavage enzymes family, which has been well studied with experimental validations. The second dataset comprises a mixture of three evolutionarily distinct

protein families that are known to be co-localized and to interact in the degradation of xyloglucan.

Dividing a homologous family

The β-keto acid cleavage enzyme family (BKACE), initially called DUF849 (Domain of Unknown Function) in the PFAM database, has been extensively characterised using high-throughput enzymatic screening (Bastard et al., 2014). Activity profiles obtained from the screening of 124 representative BKACE sequences across 20 substrates showed a strong correlation with active-site profiles. At the time of analysis, the BKACE family contained 725 sequences from 337 bacterial genomes (mostly from Pseudomonadota), as well as 3 archaeal, and 1 eukaryotic genome. This family was divided into seven main groups based on a combination of different clustering methods, including a structural classification of their active sites computed using the Active Sites Modeling and Clustering (ASMC) method (Bailly et al., 2025; de Melo-Minardi et al., 2010). The ASMC groups correlate well with the biochemical characterizations. In this study, we aimed to compare ASMC and NetSyn clusters to assess the extent of overlap between the two methods.

We retrieved the identifiers of the 725 sequences used in the ASMC analysis and submitted them to NetSyn. Of the 725 entries, 159 could not be associated with a Genbank assembly file because the corresponding UniProt entries were obsolete. Among the remaining 566 entries, 86 lacked relevant conserved genomic context (*i.e.* with a synteny score >=3). Therefore, 480 input entries were retained in the final network. The Louvain algorithm (see Mat. & Meth. section) leads to a minimum number of clusters (33), with cluster size ranging from 2 to 68 sequences (Table 1). Most of these clusters show strong agreement with the seven clusters previously defined by the ASMC method, which have been experimentally validated, with a Rand index of 0.87.

ASMC\ Louvain	G1	G2	G3	G4	G5	G6	G7	Total	max % couverture
0	5	0	1	0	60	2	0	68	G5: 88%
1	1	0	29	1	0	0	3	34	G3: 85%
2	0	0	0	0	1	0	4	5	G7: 80%
3	0	0	2	0	0	0	0	2	G3: 100%
4	2	0	0	0	0	0	0	2	G4:100%
5	0	0	0	30	0	0	1	31	G4: 97%
6	0	0	3	0	0	0	0	3	G3: 100%
7	9	0	4	1	47	0	0	61	G5: 77%
8	0	0	0	0	0	7	0	7	G6: 100%
9	0	0	0	0	0	0	10	10	G7:100%
10	0	5	0	0	0	0	2	7	G2: 71%
11	0	2	0	0	0	0	0	2	G2: 100%
12	29	2	4	4	4	2	8	53	G1: 55%
13	4	0	0	0	0	0	0	4	G1: 100%
14	0	0	0	0	0	23	0	23	G6: 100%
15	0	0	1	1	0	0	0	2	G3 : 50%
16	0	0	2	0	0	0	0	2	G3: 100%
17	0	3	0	0	0	0	0	3	G2: 100%
18	0	0	0	2	0	0	0	2	G4: 100%
19	1	27	2	1	0	0	0	31	G2: 87%
20	1	0	1	0	0	0	0	2	G1: 50%
21	0	0	0	0	0	6	0	6	G6: 100%
22	32	0	0	1	0	0	1	34	G1: 94%
23	2	0	0	0	0	0	0	2	G1: 100%
24	0	0	0	0	0	6	0	6	G6: 100%
25	2	0	7	0	0	0	47	56	G7: 84%
26	0	0	0	2	0	0	0	2	G4: 100%
27	0	0	0	4	0	0	0	4	G4: 100%
28	0	0	0	5	0	0	0	5	G4: 100%
29	0	0	0	0	0	0	2	2	G7: 100%
30	0	0	0	0	0	5	0	5	G6: 100%
31	0	0	0	0	0	0	2	2	G7: 100%
32	0	0	0	0	0	0	2	2	G7: 100%
Total	88	39	56	52	112	51	82	480	

ASMC G1 members
ASMC G2 members
ASMC G3 members
ASMC G4 members
ASMC G5 members
ASMC G6 members
ASMC G7 members
Louvain clusters populated with 2 members
Louvain clusters populated with 3 to 7 members
9 most populated Louvain clusters

Table 1: Comparison between ASMC groups and Louvain clusters

It should be noted that 12 clusters contain only 2 proteins. These clusters are considered as singleton and are of little interest for our analysis as they do not gather sufficient contextual genomic information to highlight a biological process. Most of the sequences (391) are clustered into 9 Louvain clusters (0,1,5,7,12,14,19,22,25) with a number of sequences ranging from 23 to 68. Remarkably, 89% of these clusters correlate well with the 7 ASMC groups (named here after G1 to G7; Table 1): between 77 to 100 percent of the members of one specific NetSyn cluster belong to the corresponding ASMC group. The only exception is cluster 12 in which only 55% of the sequences belong to the ASMC group G1. Indeed, the cluster number 12 contains sequences from all the groups defined by the ASMC method (Table 1). Visualisation of the network generated by NetSyn on BKACE data is available at https://doi.org/10.5281/zenodo.13860489.
By design, NetSyn built clusters of proteins according to their genomic context, while ASMC groups them according to their active sites, which directly reflects the type of substrate they can catalyse (negative, positive, polar or hydrophobic substrates). The G7 group is the perfect example of this difference. Proteins lacking at least one residue of the catalytic active site were grouped together in the ASMC G7 group and were shown not to transform any keto-acids during enzymatic screening (Bastard et al., 2014): they have been annotated as non-BKACE enzymes. The G7 group is therefore a heterogeneous group whose functions could not be studied in our previous study (Bastard et al., 2014). Interestingly, the context of the proteins clustered by NetSyn provides sufficient clues to investigate new metabolic pathways in which G7 proteins could be involved (Table S1). The proteins from the ASMC G7 group have a large diversity of genomic context and have thus been distributed in 4 large Louvain clusters (Table 1). More than half of the enzymes found in group G7 are also found in the Louvain cluster 25, which includes a locus containing a BKACE-like enzyme, a quinone-oxidoreductase, a taurine dioxygenase, an MSF transporter, a serine hydrolase and a transcriptional regulator (Table S1). This metabolic cluster has been described in *Rhodococcus erythropolis* (Actinomycetes class) in the STRING database (Szklarczyk et al., 2023) but further experimental studies are needed to determine whether this is a metabolic pathway.

The ASMC G1 group was split into two highly populated NetSyn clusters (clusters 12 and 22; Table 1). Examination of the genomic context of cluster 12 shows that it is involved in the conversion of valine to leucine, while cluster 22 is probably involved in a pathway that remains to be discovered from annotations of neighbouring genes (Table S1). The G5 group was also divided into two highly populated NetSyn clusters (clusters 0 and 7). Their genomic contexts outline the two pathways involved in carnitine degradation (Table S1).

The advantage of NetSyn over ASMC is that it is not dependent on a molecular modelling threshold. For example, seven proteins have been classified by the ASMC method in the G3 group (for which there is no specific substrate signature), because their 3D model contains all the catalytic residues. However, according to their metabolic neighbours, these proteins have

been classified in the NetSyn cluster 25 which contains sequences mainly from the G7 ASMC group (Table 1). The same applies to four proteins grouped in G1, G3 and G4, now classified in the Louvain cluster 19: the genomic neighbourhood of these four genes shares two genes (Acetoacetate:butyrate CoA-transferase α subunit and β subunit) with other members of cluster 19 (Table S2). Another example is given by 2 proteins, the 3D models of which were of poor quality (sequence identity with the reference structure respectively equal to 23% and 19%), that were grouped together, by default, in cluster G7. However, these proteins have been more coherently linked by NetSyn to putative metabolic functions present in their conserved genomic context (Table 1): probably fatty acid biosynthesis for protein C0ZQU4 (cluster 1) and deoxyribose catabolism for protein A5V766 (cluster 22).

NetSyn is also independent of the taxonomic ranks of the clustered proteins. For instance cluster 19 contains proteins from organisms with different phylum ranks: Clostridia, Fusobacteriia, Bacteroidia, Beta-proteobacteria and Gamma-proteobacteria. All the major NetSyn clusters (i.e. 12, 1, 0, 14 and 25) have a wide range of taxonomic classes. One exception is cluster 5 (30 members) which contains only alpha proteobacteria, mainly of the order Rhodobacterales. Indeed, the genomic context conservation is more likely due to taxonomic proximity of the corresponding organisms rather than functional conservation. To avoid the bias caused by the presence of comparable, or even identical, genomic contexts of very close organisms, the user can apply the "merging on taxonomy" mode of NetSyn. Comparing the NetSyn graphs obtained before and after the merge option, allows the user to detect which cluster is populated due to the diversity of the taxonomy.

Overall NetSyn was able to find ASMC groups and to suggest even more detailed clusters. As ASMC has been shown to reflect *in vitro* activities, the agreement between these two methods indicates that NetSyn is able to cluster iso-functional enzymes. NetSyn clustering is influenced more by the composition of genomic neighbourhoods than by the taxonomic origin of the target proteins, as most of the clusters characterised here are highly taxonomically diverse. Manual curation of the NetSyn clusters based on literature also have demonstrated that putative metabolic pathways can be inferred.

Interconnectivities of families in metabolic pathways

A major advantage of NetSyn is its ability to analyse non homologous but functionally related genes. Indeed, a large proportion of genes in prokaryotes, and a smaller fraction in eukaryotes, are physically linked on chromosomes and involved in a specific biological function (Nützmann et al., 2018). The most widely studied genetic organisation is the operon (Jacob and Monod, 1961), in which genes are transcribed into a single polycistronic mRNA. Operons are often conserved across species through vertical inheritance, and are therefore relatively easy to predict.

Another example of gene organisation is the Polysaccharides Utilization Loci ((PULs), which have been described in Bacteroidetes. These loci consist of physically linked genes involved in carbohydrate metabolism, including two key proteins composing a transport system: a TonB dependent receptor (namely susC) and a carbohydrate binding protein (namely susD) (Bjursell et al., 2006). Therefore a strategy for PULs prediction is based on the detection of the susC/susD pair within a cluster of Carbohydrate Active Enzymes (CAZymes) (Terrapon et al 2015). However the SusC/SusD paradigm has been challenged by the discovery of alternative sugar transport systems associated with clusters of CAZyme (Larsbrink et al 2014; O Sheridan et al, 2016), as well as by evidence that the SusC/SusD system is not necessarily co-localised with CAZymes genes (Ficko-Blean et al 2017). Therefore, methods such as NetSyn, which highlight conserved genomic contexts involving non-homologous genes, are required to reliably detect gene clusters such as PULs.

Xyloglucans (XyGs) constitute a family of ubiquitous and abundant plant cell wall polysaccharides. They are composed of (1->4)-β-glucan backbone substituted with α(1->6)-xylosyl side chain, which can be extended with additional glycosyl residues (Fig 5 panel B) (Grondin et al 2017). Xyloglucan can be degraded by a complex gene locus in the saprophytic Gram-negative bacterium *Cellvibrio japonicus,* referred to as a XyGul (Xyloglucan Utilisation Loci). This locus is composed of three glycoside hydrolases (GHs) from the GH31, GH35 and GH95 families defined in the CAZy database (Drula et al., 2022) as well as a TonB-dependent receptor (TBDR) (Larsbrink et al 2014). This particular locus has also been detected in several *Xanthomonas* strains (Vieira et al., 2021).

This example aims to demonstrate that NetSyn can identify loci containing proteins from non-homologous families, without being limited to closely related organisms. We therefore used the three key genes (xyl31A, bgl35A and afc95A) from *Cellvibrio japonicus Ueda107* as reference genes to search for homologous proteins in the UniProt database using BLAST (default parameters) (Altschul et al., 1997). Homologous sequences were selected by choosing those with at least 30% identity and 80% coverage, resulting in a dataset of 15452 candidates sequences (2212 homologous to xyl31A, 1756 homologous to bgl35A and 11484 homologous to afc95A), from a wide range of different genomes (4268 bacteria, 27 Archaea et 842 Eukaryota.

This dataset was submitted to NetSyn. Of the 15,452 entries, 251 could not be associated with a GenBank identifier and 379 with an assembly file. In addition, no conserved genomic context was identified for 1481 entries. The final network, which finally gathers 13341 entries, is available at : https://doi.org/10.5281/zenodo.15001571.
The Louvain algorithm produced 310 clusters including 6 clusters with size ranging from 910 to 3005 sequences, 7 clusters ranging from 106 to 509 sequences, 15 clusters ranging from 20 to 89 sequences and 282 clusters ranging from 2 to 18 sequences (figure 4A).

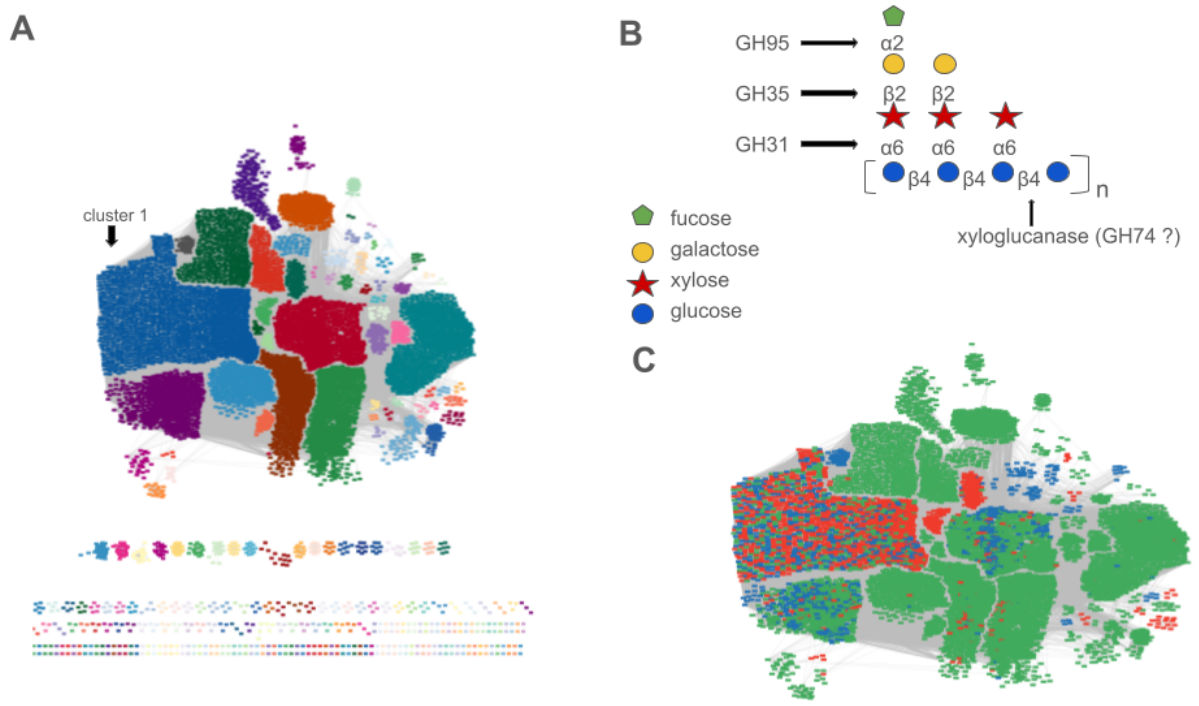


Figure 4:

Panel A: Final NetSyn network visualized with Cytoscape. In the center of the figure is shown the biggest connected components of the network. Each cluster is represented by a given color.

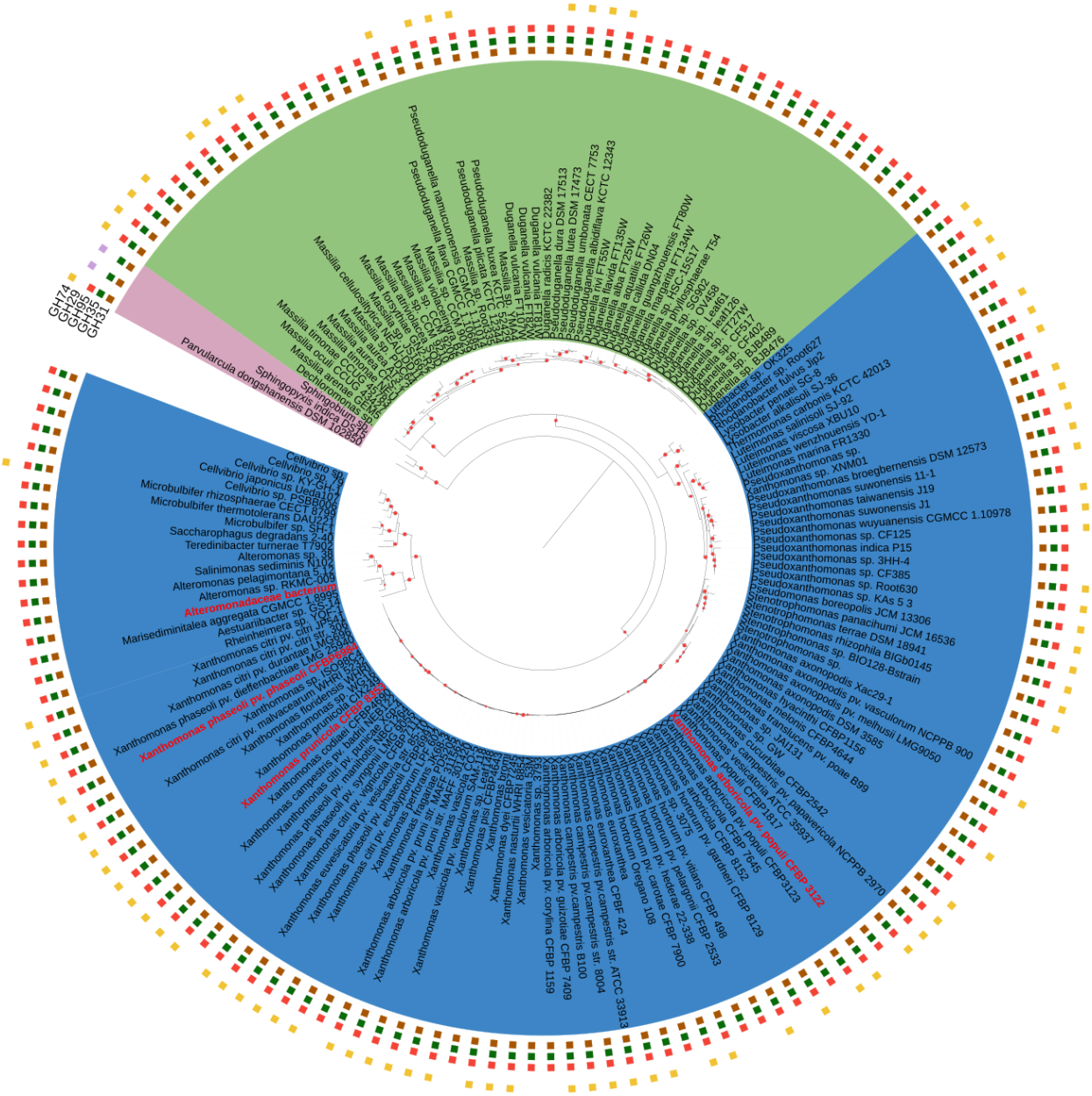
Panel B : Xyloglucan's structure and enzymes evolving in the C. japonicus Xygul (from Grondin et al (Grondin et al., 2017))

Panel C : *Zoom on part of the NetSyn network drawn by Cytoscape showing the biggest connected components. In each cluster, homologous proteins to GH31, GH35, and GH95 are respectively colored in blue, red and green.*

Only six NetSyn clusters group sequences belonging to the three GH families (clusters 1, 7, 6, 5, 2 and 10), which are most often located in different loci. Among them, only cluster 1 brings together the three GHs families (GH31, GH35 and GH95) from co-localized genes including the three reference GHs from *C. japonicus*. By investigating the loci retrieved in cluster 1, we identified 166 putative XyGuls in 162 distinct genomes belonging to different classes of Proteobacteria. Forty Xygul loci have previously been described in various *Xanthomonas* strains (Vieira et al., 2021), of which 19 were recovered by NetSyn. The remaining loci were not detected because the corresponding proteins were either not present in Uniprot, or had less than 30% identity with the sequences of *C. japonicus,* and were therefore excluded from the initial

dataset.

Interestingly, although XyGULs has been described exclusively in Gammaproteobacteria, our analysis revealed the presence of similar loci in Betaproteobacteria and Alphaproteobacteria. A more in-depth analysis of the XyGULs identified in Alphaproteobacteria is required, as in two of them, the exo-α-1,2-L-fucosidase (EC 3.2.1.63) belonging to the GH95 family is absent, and appears to be replaced by a GH29 enzyme, a family also contains exo-alpha-1,2-L-fucosidases (Wu et al., 2023). It is worth noting that the locus described in *Cellvibrio japonicus* lacks xyloglucanase activity (EC 3.2.1.151), which is found in the GH74 family, which is essential for the degradation of the xyloglucan backbone (Larsbrink et al., 2014; Vieira et al., 2021). Notably, among the 162 genomes identified in our analysis, 88 contain a XyGUL including a GH74 *(Xanthomonas, Massilia, Pseudoxanthomonas and Pseudoduganella)* (Fig 5, Table S3). This supports that we have detected loci involved in xyloglucan degradation.


GH74
GH29
GH95
GH35
GH31
Parvularcula dongshaneensis DSM 102889
Sphingopyxis indica DSM
Sphingobium sp.
Dechloromonas sp.
Massilia arenae
Massilia oculii CCUG
Massilia timonae CCUG
Massilia aurea CCUG
Massilia sp. CCUG
Massilia cellulosilytica
Massilia sp. RIVM
Massilia sp. Root
Massilia sp. CCUG
Massilia sp. CCUG
Massilia sp. CCUG

(ABC sugar transporters, ABC xylose transporters) and enzymes (D-xylose-1-dehydrogenase) were found within the putative XyGUL loci previously identified (Stephens et al., 2007). These results indicate that the locus detected in the 160 organisms, gathered in NetSyn cluster 1, is dedicated to xyloglucan degradation. Indeed, organisms in which a xyloglucan degradation locus has been predicted by NetSyn belong to different classes: Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria (Fig 5). This observation suggests that the conserved synteny within this Netsyn cluster is due more to a functional conservation than to taxonomic conservation.

The use of Netsyn enabled the detection of the XyGUL locus in many different phyla, including Alphaproteobacteria (Table 2). Finally, we show that NetSyn can group together enzymes acting in synergy into the same degrading system. However, experimental characterisation is still required to validate these predictions.

NetSyn limitation and comparison with other tools

One limitation of NetSyn is that it may be biased by the selection of sequences from organisms belonging to a narrow taxonomic range (i.e. organisms from the same genus). Indeed, a lack of genomic diversity in the input dataset may lead to the detection of synteny driven primarily by phylogenetic relatedness rather than functional conservation. To overcome this problem, we propose reducing the redundancy by merging nodes in NetSyn clusters sharing the same taxonomic rank or metadata. This approach reduces the importance of clusters grouping sequences from very closely related organisms. Another current limitation concerns the number of sequences that NetSyn can handle: beyond 10,000 sequences, graphML files may be too large to be opened by Gephi or Cytoscape, and the generated HTML file may exceed browser size limits. Furthermore, although NetSyn can process proteins from eukaryotic organisms, it was primarily designed for prokaryotic protein families and may therefore be affected by features specific to eukaryotic genomes, such as large tandem array genes.

Several tools can perform synteny analysis (e.g., Syntenet, EFI, GCsnap and CAGECAT), but direct comparison with NetSyn is challenging because these tools differ substantially in their objectives and methodologies. For instance, NetSyn was developed to explore prokaryotic protein families in order to uncover functional relationships in distantly related bacteria, whereas Syntenet focuses on whole eukaryotic genomes comparisons, notably to support phylogenetic reconstruction(Almeida-Silva et al., 2023; Guo et al., 2025; Li et al., 2025; Tao, 2025; Zhao and Schranz, 2019, 2017).
EFI groups proteins based on sequence similarity rather than genomic context conservation. Although this approach is robust, it relies on Sequence Similarity Network (SSN) driven solely

by sequence homology, without considering conservation of genomic context within the studied families. Consequently, EFI does not support analyses that combine proteins of distinct evolutionary origins, unlike NetSyn (Oberg et al., 2023).

GCsnap is methodologically similar to NetSyn, as it retrieves and compares the context of input sequences. However, NetSyn goes further, by grouping protein sequences according to the similarity of their genomic context (Pereira, 2021).

Finally, CAGECAT enables the search for homologous gene clusters in the NCBI database, but this approach requires prior knowledge of the gene cluster composition. In contrast, NetSyn can identify conserved genomic contexts without any prior information about their composition (van den Belt et al., 2023) (cf. table 2).

	NetSyn	syntenet (R package)/ SynNet-Pipeline (bash and R script)	EFI-GNT	GCsnap	CAGECAT
Purpose	Partition prokaryotic protein families based on synteny conservation	Infer synteny networks from whole-genome protein sequence data	Analyse genomic contexts of protein families from sequence similarity network (SSN) clusters	Compare genomic contexts from a list of proteins	Search for homologous gene clusters
Input	list of proteins from UniProtKB or GenBank/ENA files	complete genomes in GFF and CDS/protein in FASTA formats	EFI-EST sequence similarity network	list of proteins from NCBI Entrez, UniprotKB, UniRef, GeneID or Ensembl	GenPept protein accessions, GenBank gene clusters, Pfam identifiers, CDS/proteins fasta file or GenBank file with regions of interest
Output	Network of proteins clustered by synteny conservation	Network of genes clustered by syntenic relationships and phylogenomic profiles	Network of SSN clusters linked to PFAM domains	Genomic map showing synteny conservation	Gene presence/absence matrix and genomic map of synteny conservation

Output format	tsv files, interactive html graph and graphML file	text files, tables and graphical outputs	XGMML files and SQLite databases	text files, interactive HTML file	Interactive HTML file
Type of pipeline	python pipeline; fully automatic	R and bash scripts. Genome file download and protein family member selection are manual while homology search, synteny computation, network construction and infomap clustering are automated	web server	python pipeline, genome Download (automatic or manual) and homology search (automatic)	web server

Table 2: Comparison of tools for exploring genomic context

Conclusion

We presented NetSyn, a novel tool that links prokaryotic protein sequences within a network based on the conservation of their genomic context. It fills the gap left by current methods that rely solely on sequence similarity or active site modelling, such as ASMC, for identifying isofunctional protein groups regardless of the taxonomic rank of the source organisms.

By grouping sequences according to the conservation of their genomic context, NetSyn appears to be a valuable tool for supporting the functional annotation of protein sequences with unknown function. Application of NetSyn to two protein datasets demonstrates its ability to gather enzymes acting in the same degrading system, even in the absence of sequence similarity. Their grouping is in fact due the conservation of their genomic context which reflects underlying functional conservation. The integration of both homologous and non-homologous protein sequences as input is therefore essential when studying genomic structures such as operons, in which multiple genes involved in a specific cellular function (e.g., metabolic pathways,

regulatory processes, etc.) are co-localized within the same locus. NetSyn is therefore an effective tool for highlighting groups of non homologous proteins that participate in a common biological system, such as a metabolic pathway. Furthermore, we show that new putative enzymatic activities can be inferred from the annotation associated with genes in the genomic context of the target genes.

NetSyn is capable of constructing a network in which the relationship between two proteins depends only on the number of homologous genes within their respective genomic contexts. It can therefore be used either to explore the functional diversity of enzyme families or to identify conserved loci containing genes from different evolutionary origins involving in the same biological pathway (i.e., the same degradation/ biosynthesis process). NetSyn thus addresses the challenging task of predicting novel enzymatic activities. NetSyn is freely available at https://github.com/labgem/netsyn

References

Almeida-Silva, F., Zhao, T., Ullrich, K.K., Schranz, M.E., Van de Peer, Y., 2023. syntenet: an R/Bioconductor package for the inference and analysis of synteny networks. Bioinforma. Oxf. Engl. 39, btac806. https://doi.org/10.1093/bioinformatics/btac806

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Bailly, T., Elisée, E., Vallenet, D., 2025. ASMC: investigating the amino acid diversity of enzyme active sites. Bioinformatics 41, btaf307. https://doi.org/10.1093/bioinformatics/btaf307

Bastard, K., Smith, A.A.T., Vergne-Vaxelaire, C., Perret, A., Zaparucha, A., De Melo-Minardi, R., Mariage, A., Boutard, M., Debard, A., Lechaplais, C., Pelle, C., Pellouin, V., Perchat, N., Petit, J.-L., Kreimeyer, A., Medigue, C., Weissenbach, J., Artiguenave, F., De Berardinis, V., Vallenet, D., Salanoubat, M., 2014. Revealing the hidden functional diversity of an enzyme family. Nat. Chem. Biol. 10, 42–49. https://doi.org/10.1038/nchembio.1387

Bjursell, M.K., Martens, E.C., Gordon, J.I., 2006. Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. J. Biol. Chem. 281, 36269–36279. https://doi.org/10.1074/jbc.M606509200

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., Eisenberg, D., 2004. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 5, R35. https://doi.org/10.1186/gb-2004-5-5-r35

Boyer, F., Morgat, A., Labarre, L., Pothier, J., Viari, A., 2005. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. Bioinforma. Oxf. Engl. 21, 4209–4215. https://doi.org/10.1093/bioinformatics/bti711

Brouwer, R.W.W., Kuipers, O.P., van Hijum, S.A.F.T., 2008. The relative value of operon predictions. Brief. Bioinform. 9, 367–375. https://doi.org/10.1093/bib/bbn019

Chen, Y., Mao, F., Li, G., Xu, Y., 2011. Genome-wide discovery of missing genes in biological pathways of prokaryotes. BMC Bioinformatics 12 Suppl 1, S1.

https://doi.org/10.1186/1471-2105-12-S1-S1
Cozzetto, D., Buchan, D.W.A., Bryson, K., Jones, D.T., 2013. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics 14 Suppl 3, S1. https://doi.org/10.1186/1471-2105-14-S3-S1
de Melo-Minardi, R.C., Bastard, K., Artiguenave, F., 2010. Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinforma. Oxf. Engl. 26, 3075–3082. https://doi.org/10.1093/bioinformatics/btq595
Dongen, S.M. van, 2000. Graph clustering by flow simulation [WWW Document]. URL http://localhost/handle/1874/848 (accessed 11.17.20).
Drula, E., Garron, M.-L., Dogan, S., Lombard, V., Henrissat, B., Terrapon, N., 2022. The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res. 50, D571–D577. https://doi.org/10.1093/nar/gkab1045
Ferrer, L., Dale, J.M., Karp, P.D., 2010. A systematic study of genome context methods: calibration, normalization and combination. BMC Bioinformatics 11, 493. https://doi.org/10.1186/1471-2105-11-493
Galperin, M.Y., Koonin, E.V., 2010. From complete genome sequence to “complete” understanding? Trends Biotechnol. 28, 398–406. https://doi.org/10.1016/j.tibtech.2010.05.006
Gilchrist, C.L.M., Booth, T.J., Van Wersch, B., Van Grieken, L., Medema, M.H., Chooi, Y.-H., 2020. cblaster: a remote search tool for rapid identification and visualisation of homologous gene clusters. https://doi.org/10.1101/2020.11.08.370601
Goudey, B., Geard, N., Verspoor, K., Zobel, J., 2022. Propagation, detection and correction of errors using the sequence database network. Brief. Bioinform. 23, bbac416. https://doi.org/10.1093/bib/bbac416
Grondin, J.M., Tamura, K., Déjean, G., Abbott, D.W., Brumer, H., 2017. Polysaccharide Utilization Loci: Fueling Microbial Communities. J. Bacteriol. 199, e00860-16. https://doi.org/10.1128/JB.00860-16
Guo, B.-C., Zhang, Y.-R., Liu, Z.-G., Li, X.-C., Yu, Z., Ping, B.-Y., Sun, Y.-Q., van den Burg, H., Ma, F.-W., Zhao, T., 2025. Deciphering Plant NLR Genomic Evolution: Synteny-Informed Classification Unveils Insights into TNL Gene Loss. Mol. Biol. Evol. 42, msaf015. https://doi.org/10.1093/molbev/msaf015
Huynen, M.A., Snel, B., 2000. Gene and context: integrative approaches to genome analysis. Adv. Protein Chem. 54, 345–379. https://doi.org/10.1016/s0065-3233(00)54010-8
Jacob, F., Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356. https://doi.org/10.1016/s0022-2836(61)80072-7
Janga, S.C., Collado-Vides, J., Moreno-Hagelsieb, G., 2005. Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. Nucleic Acids Res. 33, 2521–2530. https://doi.org/10.1093/nar/gki545
Jung, J., Lee, H.K., Yi, G., 2014. A novel method for functional annotation prediction based on combination of classification methods. ScientificWorldJournal 2014, 542824. https://doi.org/10.1155/2014/542824
Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., Church, G.M., 2006. Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics 7, 177. https://doi.org/10.1186/1471-2105-7-177
Larsbrink, J., Thompson, A.J., Lundqvist, M., Gardner, J.G., Davies, G.J., Brumer, H., 2014. A complex gene locus enables xyloglucan utilization in the model saprophyte Cellvibrio japonicus. Mol. Microbiol. 94, 418–433. https://doi.org/10.1111/mmi.12776
Lee, J., Hong, W.-Y., Cho, M., Sim, M., Lee, D., Ko, Y., Kim, J., 2016. Synteny Portal: a web-based application portal for synteny block analysis. Nucleic Acids Res. 44, W35-40. https://doi.org/10.1093/nar/gkw310

Li, H., Li, Jiale, Li, X., Li, Jialin, Chen, D., Zhang, Y., Yu, Q., Yang, F., Liu, Y., Dai, W., Sun, Y., Li, P., Schranz, M.E., Ma, F., Zhao, T., 2025. Genomic investigation of plant secondary metabolism: insights from synteny network analysis of oxidosqualene cyclase flanking genes. New Phytol. 245, 2150–2169. https://doi.org/10.1111/nph.20357

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751–753. https://doi.org/10.1126/science.285.5428.751

McClean, P.E., Mamidi, S., McConnell, M., Chikara, S., Lee, R., 2010. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. BMC Genomics 11, 184. https://doi.org/10.1186/1471-2164-11-184

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419. https://doi.org/10.1093/nar/gkaa913

Nützmann, H.-W., Scazzocchio, C., Osbourn, A., 2018. Metabolic Gene Clusters in Eukaryotes. Annu. Rev. Genet. 52, 159–183. https://doi.org/10.1146/annurev-genet-120417-031237

Oberg, N., Zallot, R., Gerlt, J.A., 2023. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. J. Mol. Biol. 435, 168018. https://doi.org/10.1016/j.jmb.2023.168018

Ofer, D., Brandes, N., Linial, M., 2021. The language of proteins: NLP, machine learning & protein sequences. Comput. Struct. Biotechnol. J. 19, 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. 1, 93–108.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A. 96, 4285–4288. https://doi.org/10.1073/pnas.96.8.4285

Peng, W., Wang, J., Cai, J., Chen, L., Li, M., Wu, F.-X., 2014. Improving protein function prediction using domain and protein complexes in PPI networks. BMC Syst. Biol. 8, 35. https://doi.org/10.1186/1752-0509-8-35

Pereira, J., 2021. GCsnap: Interactive Snapshots for the Comparison of Protein-Coding Genomic Contexts. J. Mol. Biol., Computation Resources for Molecular Biology 433, 166943. https://doi.org/10.1016/j.jmb.2021.166943

Pons, P., Latapy, M., 2005. Computing Communities in Large Networks Using Random Walks, in: Yolum, pInar, Güngör, T., Gürgen, F., Özturan, C. (Eds.), Computer and Information Sciences - ISCIS 2005, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 284–293. https://doi.org/10.1007/11569596_31

Rembeza, E., Engqvist, M.K.M., 2021. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. PLoS Comput. Biol. 17, e1009446. https://doi.org/10.1371/journal.pcbi.1009446

Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., Koonin, E.V., 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. 30, 2212–2223. https://doi.org/10.1093/nar/30.10.2212

Rosvall, M., Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. U. S. A. 105, 1118–1123. https://doi.org/10.1073/pnas.0706851105

Santos, E.L. de los, Challis, G.L., 2017. clusterTools: proximity searches for functional elements to identify putative biosynthetic gene clusters. https://doi.org/10.1101/119214

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. https://doi.org/10.1371/journal.pcbi.1000605

Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I., 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database J. Biol. Databases Curation 2020, baaa062. https://doi.org/10.1093/database/baaa062

Smith, A.A.T., Belda, E., Viari, A., Medigue, C., Vallenet, D., 2012. The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Comput. Biol. 8, e1002540. https://doi.org/10.1371/journal.pcbi.1002540

Sorokina, M., Stam, M., Médigue, C., Lespinet, O., Vallenet, D., 2014. Profiling the orphan enzymes. Biol. Direct 9, 10. https://doi.org/10.1186/1745-6150-9-10

Steinegger, M., Salzberg, S.L., 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol. 21, 115. https://doi.org/10.1186/s13059-020-02023-1

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028. https://doi.org/10.1038/nbt.3988

Stephens, C., Christen, B., Fuchs, T., Sundaram, V., Watanabe, K., Jenal, U., 2007. Genetic Analysis of a Novel Pathway for d-Xylose Metabolism in Caulobacter crescentus. J. Bacteriol. 189, 2181–2185. https://doi.org/10.1128/JB.01438-06

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., Bork, P., Jensen, L.J., von Mering, C., 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 51, D638–D646. https://doi.org/10.1093/nar/gkac1000

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C., 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 45, D362–D368. https://doi.org/10.1093/nar/gkw937

Tamames, J., 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 2, RESEARCH0020. https://doi.org/10.1186/gb-2001-2-6-research0020

Tao, 2025. zhaotao1987/SynNet-Pipeline.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., Médigue, C., 2006. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 34, 53–65. https://doi.org/10.1093/nar/gkj406

van den Belt, M., Gilchrist, C., Booth, T.J., Chooi, Y.-H., Medema, M.H., Alanjary, M., 2023. CAGECAT: The CompArative GEne Cluster Analysis Toolbox for rapid search and visualisation of homologous gene clusters. BMC Bioinformatics 24, 181. https://doi.org/10.1186/s12859-023-05311-2

Vieira, P.S., Bonfim, I.M., Araujo, E.A., Melo, R.R., Lima, A.R., Fessel, M.R., Paixão, D.A.A., Persinoti, G.F., Rocco, S.A., Lima, T.B., Pirolla, R.A.S., Morais, M.A.B., Correa, J.B.L., Zanphorlin, L.M., Diogo, J.A., Lima, E.A., Grandis, A., Buckeridge, M.S., Gozzo, F.C., Benedetti, C.E., Polikarpov, I., Giuseppe, P.O., Murakami, M.T., 2021. Xyloglucan processing machinery in Xanthomonas pathogens and its role in the transcriptional activation of virulence factors. Nat. Commun. 12, 4049. https://doi.org/10.1038/s41467-021-24277-4

Wu, H., Owen, C.D., Juge, N., 2023. Structure and function of microbial α-l-fucosidases: a mini review. Essays Biochem. 67, 397–412. https://doi.org/10.1042/EBC20220158

Yamanishi, Y., Mihara, H., Osaki, M., Muramatsu, H., Esaki, N., Sato, T., Hizukuri, Y., Goto, S., Kanehisa, M., 2007. Prediction of missing enzyme genes in a bacterial metabolic

network. Reconstruction of the lysine-degradation pathway of Pseudomonas aeruginosa. FEBS J. 274, 2262–2273. https://doi.org/10.1111/j.1742-4658.2007.05763.x

Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., Gupta, D., Gupta, V., Haseeb, M., Ihsan, M., Ivanov, E., Jayathilaka, S., Kadhirvelu, V.B., Kumar, M., Lathi, A., Leinonen, R., McKinnon, J., Meszaros, L., O'Cathail, C., Ouma, D., Paupério, J., Pesant, S., Rahman, N., Rinck, G., Selvakumar, S., Suman, S., Sunthornyotin, Y., Ventouratou, M., Vijayaraja, S., Waheed, Z., Woollard, P., Zyoud, A., Burdett, T., Cochrane, G., 2024. The European Nucleotide Archive in 2023. Nucleic Acids Res. 52, D92–D97. https://doi.org/10.1093/nar/gkad1067

Zallot, R., Oberg, N., Gerlt, J.A., 2019. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry 58, 4169–4182. https://doi.org/10.1021/acs.biochem.9b00735

Zhao, B., Hu, S., Li, X., Zhang, F., Tian, Q., Ni, W., 2016. An efficient method for protein function annotation based on multilayer protein networks. Hum. Genomics 10, 33. https://doi.org/10.1186/s40246-016-0087-x

Zhao, T., Schranz, M.E., 2019. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proc. Natl. Acad. Sci. U. S. A. 116, 2165–2174. https://doi.org/10.1073/pnas.1801757116

Zhao, T., Schranz, M.E., 2017. Network approaches for plant phylogenomic synteny analysis. Curr. Opin. Plant Biol., 36 Genome studies and molecular genetics 36, 129–134. https://doi.org/10.1016/j.pbi.2017.03.001

Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., Kasif, S., 2002. Computational identification of operons in microbial genomes. Genome Res. 12, 1221–1230. https://doi.org/10.1101/gr.200602