NetSyn: genomic context exploration of protein families

Mark Stam, Jordan Langlois, Céline Chevalier, Guillaume Reboul, Karine Bastard, Claudine Médigue, David Vallenet

NetSyn: genomic context exploration of protein families

Mark Stam[1], Jordan Langlois[1], Céline Chevalier[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
CEA - Genoscope, 2 rue Gaston Crémieux CP 5706, 91057 Evry Cedex, France
Email address: mstam@genoscope.cns.fr

Abstract

Background: The growing availability of large genomic datasets presents an opportunity to discover novel metabolic pathways and enzymatic reactions profitable for industrial or synthetic biological applications. Efforts to identify new enzyme functions in this substantial number of sequences cannot be achieved without the help of bioinformatics tools and the development of new strategies. The classical way to assign a function to a gene uses sequence similarity. However, another way is to mine databases to identify conserved gene clusters (i.e. syntenies) as, in prokaryotic genomes, genes involved in the same pathway are frequently encoded in a single locus with an operonic organisation. This Genomic Context (GC) conservation is considered as a reliable indicator of functional relationships, and thus is a promising approach to improve the gene function prediction.
Methods. Here we present NetSyn (Network Synteny), a tool, which aims to cluster protein sequences according to the similarity of their genomic context rather than their sequence similarity. Starting from a set of protein sequences of interest, NetSyn retrieves neighbouring genes from the corresponding genomes as well as their protein sequence. Homologous protein families are then computed to measure synteny conservation between each pair of input sequences using a GC score. A network is then created where nodes represent the input proteins and edges the fact that two proteins share a common GC. The weight of the edges corresponds to the synteny conservation score. The network is then partitioned into clusters of proteins sharing a high degree of synteny conservation.
Results. As a proof of concept, we used NetSyn on two different datasets. The first one is made of homologous sequences of an enzyme family (the BKACE family, previously named DUF849)

to divide it into sub-families of specific activities. NetSyn was able to go further by providing additional subfamilies in addition to those previously published. The second dataset corresponds to a set of non-homologous proteins consisting of different Glycosyl Hydrolases (GH) with the aim of interconnecting them and finding conserved operon-like genomic structures. NetSyn was able to detect the locus of Cellvibrio japonicus for the degradation of xyloglucan. It contains three non-homologous GH and was found conserved in fourteen bacterial genomes.
Discussion. NetSyn is able to cluster proteins according to their genomic context which is a way to make functional links between proteins without taking into count their sequence similarity only. We showed that NetSyn is efficient in exploring large protein families to define iso-functional groups. It can also highlight functional interactions between proteins from different families and predicts new conserved genomic structures that have not yet been experimentally characterised. NetSyn can also be useful in pinpointing mis-annotations that have been propagated in databases and in suggesting annotations on proteins currently annotated as "unknown". NetSyn is freely available at https://github.com/labgem/netsyn.

Introduction

The advance of genome sequencing technology has created an ever increasing gap between protein sequences and annotations (Galperin and Koonin, 2010). It is currently estimated that 25% of the families defined in the PFAM database are annotated as "unknown function" (Mudgal et al., 2015). Moreover, this issue is worsened by current automatic annotation procedures that still have a high false positive rate. This is especially true for methods that rely only on sequence similarity which lead to some protein families having a level of mis-annotation as high as 80% (Schnoes et al., 2009). To overcome the lack or error annotation, innovative strategies use 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), Protein Protein Interaction (PPI) network (Cozzetto et al., 2013; Peng et al., 2014; Zhao et al., 2016).

A way to circumvent the challenge of the gene function annotation, is to compare the genomic context (GC) of genes which is considered as 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). GC methods can be classified into four categories (Ferrer et al., 2010): gene neighbour, gene cluster, gene fusion (or Rosetta Stone) andphylogenetic profile and . Gene cluster (or conserved synteny) method is defined as the conservation of chromosomal proximity between genes (Overbeek et al., 1999; Tamames, 2001). It is based on conserved genes through different organisms and computes distances, in number of bases or genes, between two adjacent genes that are transcribed in the same direction (Overbeek et al., 1999). Such distances are

considered as a good predictor for co-regulated groups of genes like in operon structure (Brouwer et al., 2008) or polysaccharide utilisation locis (PUL) (Bjursell et al., 2006). The Phylogenetic profile methods (Pellegrini et al., 1999) create a presence/absence matrix of homologous genes through different genomes. The length of the matrix depends on the number of considered genomes. It assumes that genes having a similar phylogenetic profile are more likely to share the same fonction. The gene fusion (Marcotte et al., 1999) (or rosetta) method is based on genes fusion events. The idea is that two genes can be fused into some genomes. Pairs of proteins linked by such a fusion relationship are more likely to be functionally linked and occur in the same pathway. The gene neighbour methods (Bowers et al., 2004) calculate the distance between two genes in different genomes. If one considers two genes from a genome, the distance between them is the number of genes between them plus one. This distance can be computed into different reference genomes. If the distance increases that means that the two genes are not linked. Two linked genes are more likely to be functionally in interaction. But in order to avoid a taxonomy bias, one must choose carefully the studied genomes which must not be too close in the taxonomy (*i.e.* organisms from the same genus).

Other efforts have been carried out to incorporate information from gene context to integrative approach. For instance, the STRING database (Szklarczyk et al., 2017) stores all the known and predicted protein-protein interactions from different sources. These interactions can be physical (*i.e.* heterodimer) or indirect (*e.g.* gene regulation, signal relay mechanism) and can be represented by a graph where nodes correspond to the proteins and the edges to the interaction. This network highlights the interactions between different protein families. Another example is the Enzyme Function Initiative (EFI) (Zallot et al., 2019) which has developed a method based on the Genomic Neighborhood Network (GNN). Starting with a list of query sequences, a Sequences Similarity Network (SSN) is created, in which groups of highly connected nodes are considered as a cluster. Then, the genomic context is explored by recovering neighbour genes from the European Nucleotide Archive (ENA) (Kanz et al., 2005).Finally, neighbour genes are considered related if they belong to the same PFAM family (Finn et al., 2014). As output, EFI gives a Genome Neighbourhood Diagram (GNDs) for visual representation of the neighbourhoods for the sequences in each SSN cluster. In another study, gene context clustering had been integrated in a mega-clustering approach, which also included active site modeling, sequence similarity and phylogeny. Applied to a family of unknown function PFAM DUF849, forteen potential new activities had been predicted and experimentally validated subsequently (Bastard et al., 2014).

None of the methods cited above have developed a sequence network that reflects only the GC conservation while avoiding the sequence similarity. A sequence similarity steps limit the input to only one homologous family at once, and make them unable to handle a set of non homologous proteins. Mixing non-homologous proteins as input is necessary when one wants to

study genomic structures like operon, in which several genes involved in a specific cellular function (*i.e.*, metabolic pathway, regulation, etc..) are co-localized in the same locus. With a method based only on the synteny conservation, genes belonging to the same structure will be gathered into the same cluster without having any sequence similarity.

Here, we present NetSyn, for Network Synteny, a tool allowing to link sequences based on the conservation of their genomic context in a network. NetSyn also offers several functionalities to explore the synteny network such as clustering the sequences, visualizing the genomic context of each sequence or mapping data from diverse resources. Our method quantifies genomic context conservation of proteins taken from different genomes. The advantage of NetSyn over others methods, is the fact that the sequence comparison is based not on sequence similarity but only on genomic context conservation. This important originality allows the genomic comparison of both a set of homologous sequences like a whole sequence family, as well as a set of non-homologous sequences like several proteomes. NetSyn was tested on two kinds of dataset. One set contains sequences from an homologous family. NetSyn was able to create iso-functional sub-families and pinpoint putative new pathways. The second set contains sequences from three non-homologous families, involved in the same degradation pathway. NetSyn was able to retrieve this known pathway in different organisms. Widely, the NetSyn principle is based on the idea that proteins sharing a similar genomic context are more likely to share similar functions or being involved in the same metabolic pathway. Consequently, our method can be extended to various sets of proteins.

Materials & Methods

Workflow and principles of NetSyn

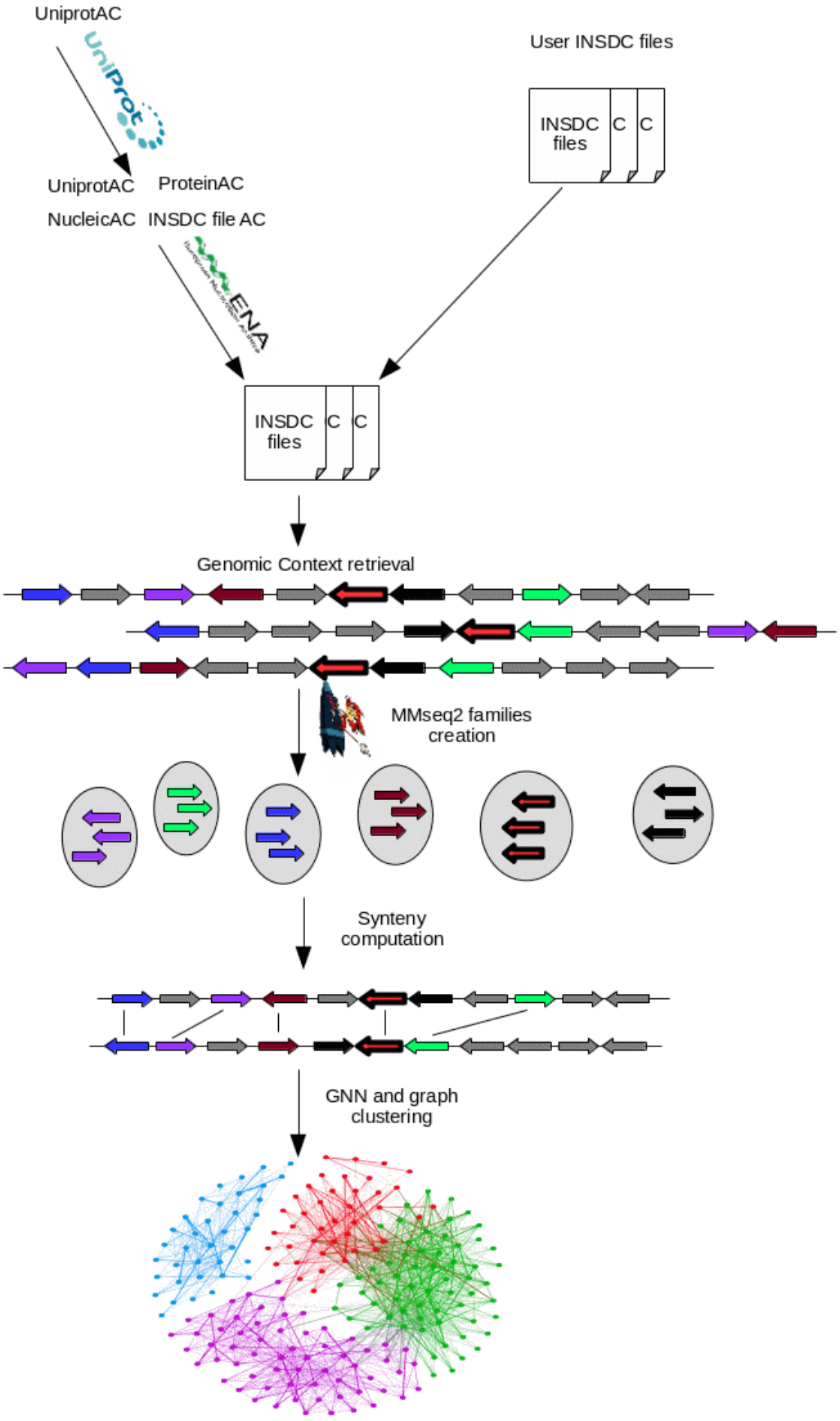
NetSyn is a workflow developed in Python 3 and is divided into four steps: i) downloading the data, ii) extracting the genomic context, iii) computing the synteny conservation, iiii) computing a synteny network. NetSyn accepts as input a list of UniProt accession numbers (hereafter called target proteins). The workflow makes requests to the UniProt rest API to retrieve the corresponding nucleic accession number and then download the genome assembly file (hereafter called the genome file) from the European Nucleotide Archive (ENA) web site. Genomes files are used to extract the GC for each target protein (Fig. S1). But one can use his own data that are not stored in UniProt and ENA. In that case, NetSyn takes as input a correspondence file (Fig. S2), which is in a tabulation-separated value format with 5 mandatory fields:

- protein_AC: the protein accession number in the genome file
- protein_AC_field: the name of the field (qualifier) from which the protein accession number can

be found in the genome file
- nucleic_AC: the accession number of the genome file
- nucleic_File_Format: the format of the genome file (EMBL or Genbank format)
- nucleic_File_Path: the path to the locally stored genome file.

In addition to these files, NetSyn can accept a user-supplied metadata file to give attributes to target proteins. These attributes can be displayed on the final synteny network (example is given in Fig. S3). NetSyn parses the different genome files to retrieve the sequence of the target proteins and also the protein sequence of the neighbour genes. The number of retrieved neighbour genes depends on the window size parameter. It is set to 11 by default and means that NetSyn takes into account 5 genes before and 5 genes after the target gene. NetSyn also recovers, from the genome files, the full taxonomic lineage of the organisms having the target proteins. This taxonomic information is reported on the final graph and can be used to remove the redundancy if the graph is too dense (see the paragraph “Merging nodes”). All the retrieved proteins are then clustered by MMseqs2 (Steinegger and Söding, 2017) based on their sequence identity and alignment coverage percentages. These parameters can be modified and are set to 30% of identity on 80% of coverage by default, respectively. These MMseqs2 clusters allow NetSyn to define homologous protein families that are used to compute the synteny conservation for each pair of target proteins (Fig. 1).

Figure 1



Synteny computation and scoring

The computation of the genomic context conservation (i.e. synteny conservation) between two target proteins is made by an exact graph-theoretical approach described by Boyer *et al.* (Boyer et al., 2005). For each pair of target proteins in two different genomes, namely GA and GB, the graph-theoretical approach defines two networks, namely N1 and N2 where the vertices are the genes and the edges represent the fact that two genes are physically linked (i.e. on the same chromosome) or are homologous. Then, it searches connected components which are a subset of a network where any two vertices are connected to each other by paths but have no edges with the rest of the network. Finally a Common Connected Component (CCC) is computed taking into account what we call "gaps". Here we consider a "gap" in genome A as a gene in genome A which has no homologous gene into the genome B. A CCC is composed of common vertex/edge relationships in connected components in N1 and in N2. Figure 2 (panel A) shows an example of the computation of a CCC. The genome A has a connect component composed with the following relationships: ($\{G_{A5}, G_{B5}\}$, $\{T_a, T_b\}$ $\{G_{A6}, G_{B6}\}$ $\{G_{A7}, G_{B6}\}$ $\{G_{A8}, G_{B8}\}$) and the genome B has a connect component composed with these relationships: ($\{G_{A1},G_{B2}\}$ $\{G_{A5}, G_{B5}\}$ $\{T_a, T_b\}$ $\{G_{A6}, G_{B6}\}$ $\{G_{A7}, G_{B6}\}$ $\{G_{A8}, G_{B8}\}$). The resulting CCC is composed by relationships present in the two lists: ($\{G_{A5}, G_{B5}\}$ $\{T_a, T_b\}$ $\{G_{A6}, G_{B6}\}$ $\{G_{A7}, G_{B6}\}$ $\{G_{A8}, G_{B8}\}$) (Fig. 2 panel B). The $\{G_{A1}, G_{B2}\}$ relationship is not taken into account in GA, because the number of genes between G_{A1} and T_A which is superior to the threshold (gap parameter set to 3 by default). The final CCC corresponds to the synteny considered by NetSyn.

Figure 2

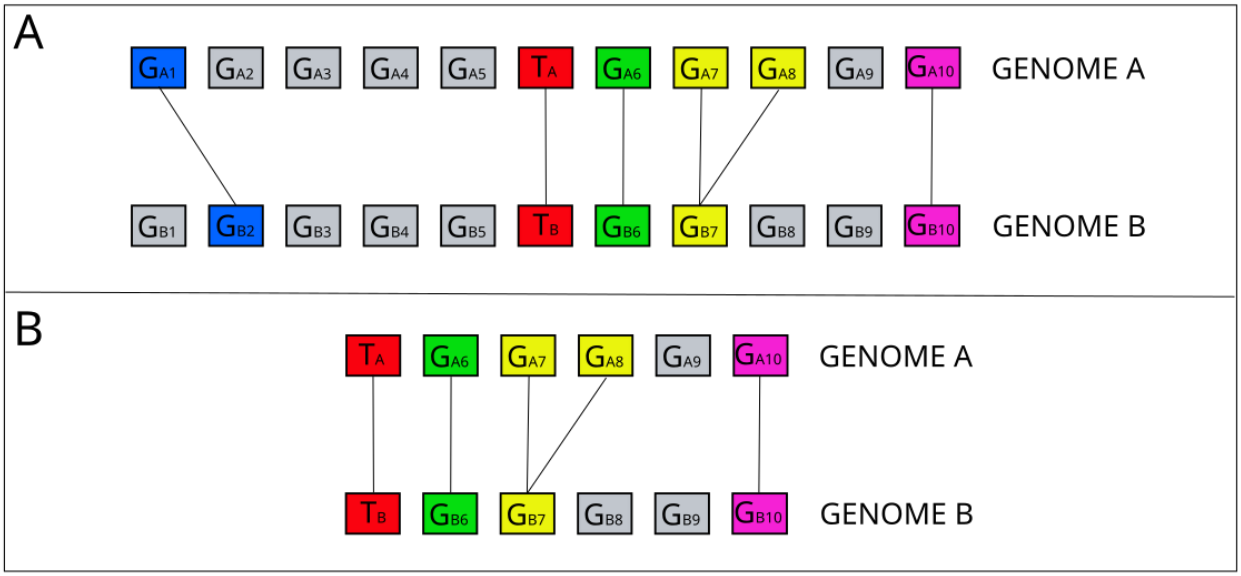


Figure 2: Example of synteny computation between two target genes.

A) Computation of a synteny between two different genomic contexts of two target genes: Ta from genome A and Tb from genome B. Homologous genes are represented by colored rectangles. In this example the window size is set to 11 and the gap parameter to 3. It means that NetSyn take into account the 5 genes before and the 5 genes after the genes target and allow a maximum of 3 genes without any homology between 2 genes with homologues in the second genome. Therefore, in this example the synteny considered by NetSyn is composed by the following relationship : Ta/Tb, G_{A6}/G_{B6}, G_{A7}/G_{B7}, G_{A7}/G_{B7}, G_{A8}/G_{B7}, G_{A10}/G_{B10}.The G_{A1}/G_{B2} relation is not considered as part of the synteny due to the gap between G_{A1} and T_A.

B) Once the synteny between two target genes, Ta and Tb is defined, the score is computed as follow : score = (gs/2)*(gs/gt)

with gs=number of genes in synteny (here gs = 9) and gt = total number of genes in the considered synteny (here gt = 12). Here the score between the two target genes T_A and T_B is Synteny Score = (9/2)*(9/12)=3,375

Once a synteny has been defined between two target proteins, a synteny score is then calculated as follow:

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

Where GS is the number of genes having an homologous gene into the opposite genome, and GT the total number of genes in the considered synteny. In Figure 2 (panel B), the considered synteny contains 9 genes in synteny (or 9 genes having an homologous gene into the opposite genome) and 12 genes in total. Therefore the synteny score is SS = (9/2) * (9/12) = 3.375.
In order to consider only significant synteny, by default, NetSyn keeps only synteny with a score equal to or higher than 3. This threshold can be modified by the user.

Synteny network and clustering

Once all syntenies and scores are computed between all the target proteins, NetSyn constructs a synteny network where nodes represent the target proteins. Two target proteins are

linked with an edge if a conserved synteny is detected between them. In order to keep only significant syntenies, NetSyn considers only syntenies with score higher than a threshold defined by the user (3 by default) for the generation of the synteny network. Target proteins that are unrelated to any others target proteins are not taken into account for synteny network construction. The weight of the edge is equal to the synteny score which makes it possible to highlight the most conserved synteny.

In order to bring out target proteins sharing the most similar genomics context, several network clustering algorithms are implemented into netsyn : MCL, Walktrap, Louvain and Infomap. The Markov Cluster algorithm (Dongen, 2000) (MCL) is an unsupervised algorithm based on simulation of flow in graphs and simulates random walks within a graph. The Walktrap algorithm (Pons and Latapy, 2005) is the only implemented algorithm which takes into account the weight of the edge. It is also based on a random walk but unlike MCL, it can handle large graphs (> 1000 vertices). Louvain (Blondel et al., 2008) is based on the comparison between the edge's density inside and outside a community (i.e. a cluster). Infomap (Rosvall and Bergstrom, 2008) is also based on a randomized walk trajectory and tries to minimize the length of the trajectory into a cluster. All the implemented algorithms can be applied on the final network. The clusters thus created are considered as the NetSyn cluster.

NetSyn network visualization

The final NetSyn's output is a network. We provide two different files format to explore and analyse the network : a graphML file which can be opened with network visualisation tools like Gephi or Cytoscape, and a HTML file, generated with the 3d-force-graph library, and which can be viewed with web browser (Fig. 3).

Figure 3

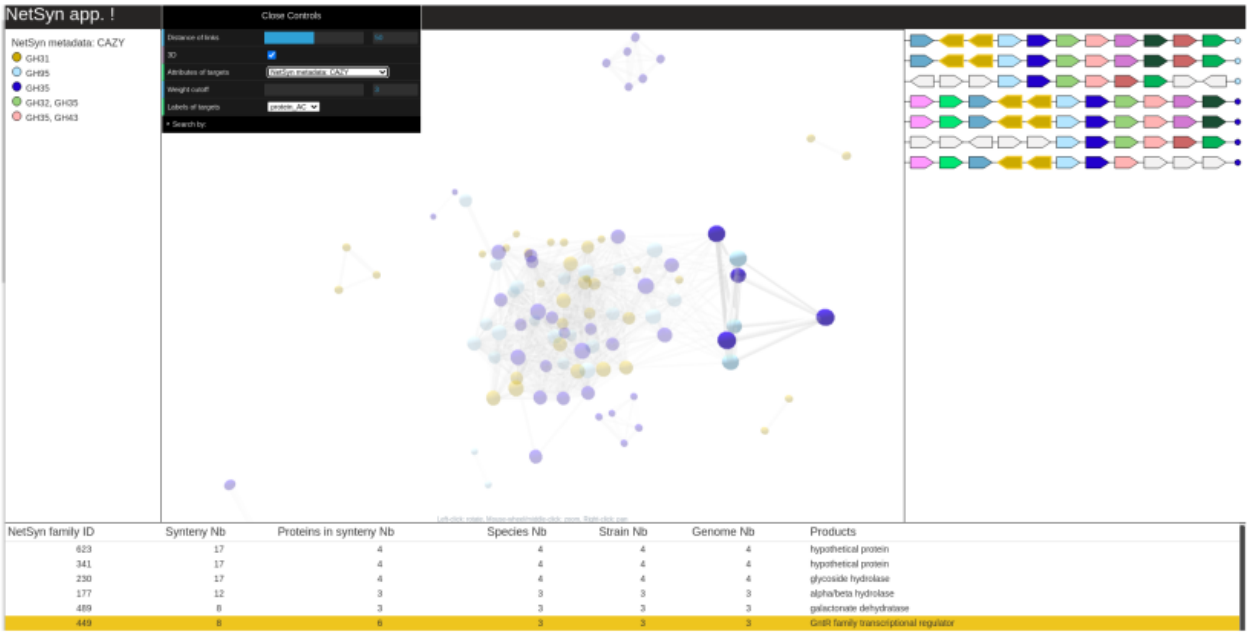


Figure 3: **example of visualization of the graph result.**

The NetSyn web interface is divided into 4 panels. The upper left panel is the legend of the network. Here the color correspond to the different CAZy families the proteins belongs to. One can click on a colored dot (here GH35) in order to select all the nodes with the same property. The upper middle panel is the final network, the color depend of the asked informations to map (here the CAZy families) : network cluster, taxnomy, metadata ... On the top, a control panel allow to select which property to display on the left panel and to search nodes with a particular label, products, EC number or MMseq ID family (see text and below). On the upper right panel, an example of a visualization of the different genomic context in a selected cluster. Each arrow represent a gene and genes having a same color are considered as homologous. The different genomic context are centered to the target protein given in input (here in yellow). On the bottom panel a description of the different MMSEQ families found into a cluster. For each MMSEQ family there are the following information: the family identification ID (NetSyn family ID), the number of times this family is in synteny between two genomes in the considered cluster (Synteny Nb field), number of proteins in this family and in this cluster (Proteins in Synteny Nb field), the number of species considered in this cluster (Species Nb field), the number of strain in this cluster (Strain Nb field), the number of genomes in this cluster (Genome Nb field), the different annotation (Products field) find for the proteins of this family (from the Product field in the genomes files).

The HTML view is divided into 4 panels. The synteny network is displayed on the upper central panel. The upper left panel is the legend of the network and displays all attributes linked to each protein target like the NetSyn cluster they belong to, the metadata provided by the user and the different taxonomic rank retrieved during the genomic file parsing step. With the control box on the top, the user can choose which attributes to display. On this panel, if a user clicks on one attribute, all nodes sharing the same attribute are highlighted on the network and a schematic view of the context appears on the upper right panel. Each context is centered on the target protein with the number of genes before and after depending on the window size parameter given

in input. In the example Fig. 3, the selected window size was 11, therefore 5 genes before and 5 genes after are shown. In this view, homologous genes are colored with the same color. If the user pass his cursor above the one the gene, a pop up appears and shows information from the NetSyn computation and the genome file: the family id (MMseqs2 family identifier), strain name, protein accession number, Uniprot accession number, gene name, locus tag, product and EC_number (Fig. S4). The lower panel gives information about the computation of homologous protein families by MMSEQ, for the selected NetSyn cluster. For each MMseqs2 family determined, it gives the number of selected proteins belonging to this family, the number of different species, strains and genomes from which the selected proteins come.

Merging nodes and network reduction

In some cases, the final network can be huge and tricky to analyse. To circumvent this challenge, we add an option to reduce the network. For that purpose, nodes identified as belonging to the same synteny cluster (several clustering methods are available) and sharing a given property will be merged into a unique node. The property can be a taxonomic rank, which is recovered by NetSyn when parsing the INSDC file, or a property given into the medatafile (Fig. S5). The merging can be applied on only one property at a time in a single NetSyn run. The called “merged nodes” are the result of the union of one NetSyn cluster and one property. Merged nodes and the associated network cannot be rebuilt after a NetSyn computation. Only one clustering method must be provided.

Results

Dividing homologous family

NetSyn was tested on the β-keto acid cleavage enzymes family (BKACE). This family, initially called DUF849 (Domain of Unknown Function) in the PFAM database, was extensively characterized using high-throughput enzymatic screening (Bastard et al., 2014). The profiles of activity issued from the enzymatic screening of 124 representatives on 20 substrates showed a high correlation with the profiles of active sites. Indeed, the 725 proteins of the BKACE family were divided into 7 main groups based on structural classification of active sites by the Active Sites Modeling and Clustering (ASMC) method (de Melo-Minardi et al., 2010). As ASMC groups show a good match with *in vitro* enzymatic activities, our aim here is to verify NetSyn clustering can identify the ASMC groups, and thus check if NetSyn can cluster iso-functional enzymes.

We submitted the same set of 725 UniProt entries to NetSyn and it took 287 seconds on

one CPU to compute. Among the 725 entries, 159 cannot be associated with an ENA file, because they were obsolete UniProt entries. Among the remaining 566 entries, 86 have no relevant conserved genomic context (*i.e.* with a synteny score >=3). Therefore 480 input entries (85% of the entries that had an corresponding ENA file) are retrieved into the final network. The clustering algorithm which gives about the same number of clusters as ASMC´s grouping was the Louvain algorithm. It gave 33 clusters with sizes varying from 2 to 68 sequences (Table 1).

Table 1: Comparison between ASMC groups and Louvain clusters

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	

	9 most populated Louvain clusters
	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

It should be noted that 12 clusters contain only 2 proteins. These clusters are considered as singleton and present low interest to analyse as they are not gathering enough context genomic information to highlight a metabolic pathway. 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 varying from 23 to 68. All these clusters show a good fit with the ASMC groups (named here after G1 to G7) meaning that between 77 to 100% of their members belong to the same ASMC group. The only exception is the Louvain cluster 12 where only 55% of its sequences belong to the group G1. Cluster 12 is the only cluster that contains sequences from all the ASMC groups (Fig. 4). The network generated by NetSyn on BKACE data is available for visualization at https://doi.org/10.5281/zenodo.5746499

Figure 4

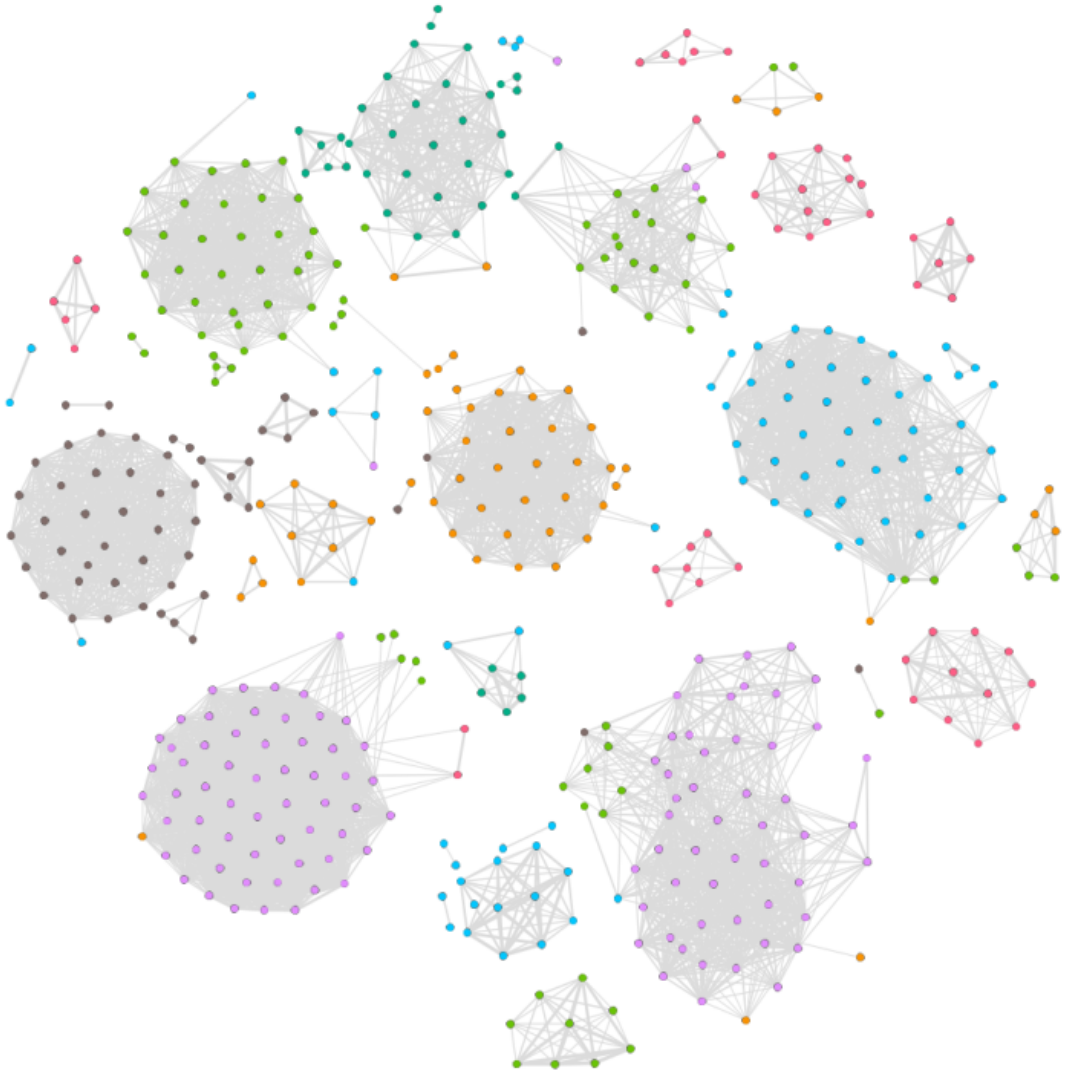


Figure 4. **BKACE's family network**

Network of the 480 proteins having a conserved genomic context in the BKACE's family (gephi output). Only the edge between nodes belonging to the same Louvain cluster are represented. The color corresponds to the 7 ASMC groups : group 1 (light green), group 2 (dark green), group 3 (orange), group 4 (brown), group 5 (pink), group 6 (red), group 7 (light blue)

By definition, NetSyn clusters proteins according to their metabolic context while ASMC gathers proteins according to their active sites, which directly reflects the type of substrate they can catalyze (negative, positive, polar or hydrophobic substrates). Group G7 is the perfect example for illustrating this difference. Proteins which lacked at least one catalytic active site residue had been grouped in the G7 (Bastard et al., 2014) and have been shown not transforming any keto-acid from the enzymatic screening. They have been annotated as non-BKACE enzymes. Thus, the G7 group is a heterogeneous group which functions could not be studied in our previous study (Bastard and Smith, 2014). However, NetSyn provides enough clues to investigate new metabolic pathways for G7 proteins. Proteins from the G7 group are distributed in 4 large Louvain clusters (Table 1) and show a large diversity of genomic context (Table S1). More than half the G7 enzymes are found in cluster Louvain 25 (Fig. S6), with a locus integrating a quine-oxidoreductase, a taurine dioxygenase, an MSF transporter, a porin, a fatty acid-CoA ligase and a transcriptional regulator (Table S1). This protein network has been described in Rhodococcus rhodochrous (actinobacteria class) (https://string-db.org/network/1429046.RR21198_4245). Other examples show that NetSyn performs better at splitting proteins according to the metabolic pathway rather than their active site. G1 was split into two highly populated NetSyn clusters (cluster 12 and 22) (Table 1). According to the genomic context, cluster 12 is involved in the conversion from valine to leucine while cluster 22 is involved in the pathway yet to be discovered, but for which clues are available (Table S1). The G5 group was also divided into high populated Louvain clusters (cluster 0 and 7). Their genomic contexts outline the two routes involved in the carnitine degradation (Table S1).

The benefit of NetSyn over ASMC is the lack of dependency on the molecular modeling threshold. 7 members previously grouped in G3 groups belong to Louvains cluster 25 (Table 1). Because the 3D model of these proteins presented all the catalytic residues, ASMC could not group them with the G7 group (i.e. at least missing one catalytic residue). Consequently, ASMC classified them in the G3 group for which there is not a particular substrate signature. However, according to their metabolic neighbors, these proteins are classified in cluster 25. Same remark occurs for four proteins classified in G1, G3 and G4 now classified in Louvain cluster 19 (Table S2). The genomic neighbouring of these four genes has in common two genes (Acetoacetate:butyrate CoA-transferase α subunit and β subunit) with other members of cluster 19 (Table S2). Contrastly, there are 8 proteins which used to belong to G7 group and now belong to cluster 12 (mostly populated with G1 group members). For instance, C0ZQU4 and A5V766 whose 3D model was of poor quality (sequence identity with the template for homology modeling at 23% and 19 %) were grouped into G7 because one catalytic residue was missing in the 3D model, whereas their metabolic function seems in a good agreement with Louvain cluster 1 and 22 (Table S3).

NetSyn is independent of the taxonomic ranks of the clustered proteins. For instance

Louvain cluster 19 contains proteins from organisms with different phylum ranks: Clostridia, Fusobacteriia, Bacteroidia, Beta-proteobacteria and Gamma-proteobacteria, which belong to different phylums (Fig. S6). All the large clusters, as cluster 12, 1, 0, 14 and 25 present a large diversity of taxonomic classes. Only the highly populated cluster 5 with 30 members contains only alpha proteobacteria, mostly with the order Rhodobacterales. All genomes from the organisms of cluster 5 are similar. By consequence, the cluster 5 is not heterogeneous and genomic information repeated a few times. The group seems highly populated but actually the genomic information is repeated a few times. To avoid this skew, it appears necessary to prepare the data by removing the closest organism from the initial set before running NetSyn. Another way of detecting this skew is to use the merging mode of NetSyn with merging on the taxonomy. By comparing the NetSyn graph before and after merging, the user can detect what cluster is populated because of the diversity of taxonomy. We have also shown by looking at all clusters and the redundant genes retrieved, it is possible to highlight new pathways as shown in Table S1.

In our results, NetSyn was able to retrieve the ASMC groups and suggest even more detailed clusters. As ASMC had been shown to reflect *in vitro* activities, the agreement between these two methods indicates that NetSyn groups iso-functional enzymes. NetSyn clustering is more influenced by the composition of genomic neighborhoods than the taxonomic origin of the target proteins as most of the clusters defined here show a large taxonomic diversity. We have also demonstrated by curating manually the NetSyn clusters with literature that putative pathways can be deduced.

Interconnectivities of families in metabolic pathways

A great advantage of NetSyn is its ability to handle genes with no homology but functionally related. Indeed, half the genes of prokaryotes and a minor fraction of genes in eukaryotes form groups of physically clustered genes on chromosomes that are related in function (Nützmann et al., 2018). The most studied gene organization is the operon (Jacob and Monod, 1961) in which genes are transcribed as single polycistronic mRNA. Operons are often conserved across species by vertical inheritance, and thus operons can be computationally predicted.

More recently an operon-like structure, grouping CAZymes (for Carbohydrate Active enZymes), transcriptional regulators and transporters, has been highlighted (Bjursell et al., 2006). In Bacteroidetes, this system is called Polysaccharide Utilization Locus (PULs) and is composed of gene clusters that encode enzyme and protein ensembles required for the saccharification of complex carbohydrates. The first described PULs was centered around a transport system

composed of two proteins : a TonB-dependent receptor, namely susC, and a carbohydrate binding protein, namely susD. Therefore, the different efforts to detect PULs were based on the research of the SusC/SusD couple followed by the inspection of the glycosyl hydrolase (GH) composition nearby (Terrapon et al., 2015). But the SusC/SusD paradigme is challenged by the discovery of different sugar systems transport associated to PULs (Larsbrink et al., 2014; O Sheridan et al., 2016) or by the fact that the SusC/SusD system is not necessary to be localized into the gene cluster (Ficko-Blean et al., 2017). Consequently, in order to detect gene clusters or PULs, a method which is freed from the search for a particular gene and which highlights the colocalization of non homologous genes is necessary. NetSyn is able to handle genes with no homology but having similar GC, like genes involved in the xyloglucan PUL.

Xyloglucans (XyG) are members of the polysaccharides family found in terrestrial plant cell walls. They are composed of a β(1→4)-glucan backbone appended with branching α(1→6) xylosyl moieties (Fig. 5 panel A) (Grondin et al., 2017). Larbrink et al. (Larsbrink et al., 2014) have demonstrated that *Cellvibrio japonicus strain Ueda107* can degrade the xyloglucan with a gene locus composed by 3 genes belonging to glycosyl hydrolase families GH31, GH35 and GH95 (genes xyl31A, bgl35A and afc95A respectively) and a transporter. Our goal is to focus on the interconnectivity between the GH proteins from different CAZy families that intervene in glycan degradation. For this purpose, we started to use the three key gene (xyl31A, bgl35A and afc95A) of *Cellvibrio japonicus strain Ueda107* as reference to search homologous gene among the complete proteomes in UniProt using Blast (default parameter) (Altschul et al., 1997)

We found 20 organisms containing homologs for the 3 key genes present in the Xygul in *Cellvibrio Japonicus* (Table S4). The complete proteome of these organisms were downloaded from UniProt and we used DBCAN (Zhang et al., 2018) to retrieve all the GHs. Then we selected all proteins annotated as GH31, GH35 or GH95 in these proteomes, which give a set of 194 UniProt entries and use this set as input for NetSyn. It took 47 second on one CPU to compute the final graph. The final network is available at https://zenodo.org/record/5705596. Only 136 input entries had a conserved genomic context and have been clusterized into 34 clusters with the Walktrap algorithm (Fig 5 panel A). Most of them are small clusters with only two or three sequences. The most populated cluster (cluster 0 in green Fig. 5 panel B) is also the only cluster where the three keys GH of the xyloglucan PUL are grouped (Fig. 5 panel C & D) including the GHs of the C. japonicus xyloglucan PUL.

Figure 5

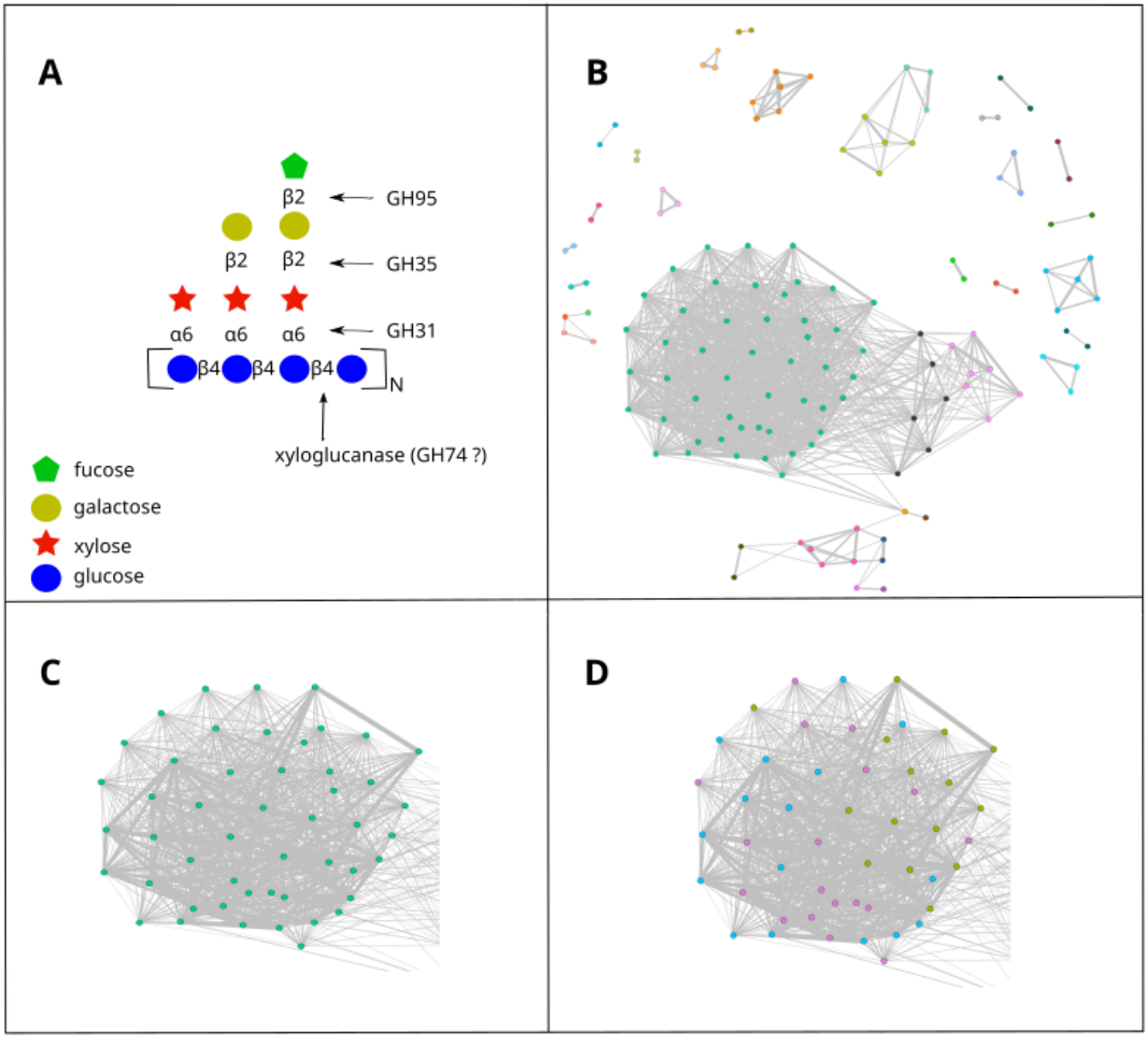


Figure 5

Panel A: structure of the xyloglucan

Panel B: Network of the GH31, GH35 and GH95 found in the 20 selected organisms. The color correspond to the different walktrap cluster comuted by NetSyn

Panel C: Cluster 0 in the network

Panel D: Cluster 0, the color correspond to the different GH family (GH31 green, GH35 pink, GH95 blue)

Then, we manually investigate the genomic context around the other GH31, GH35 and GH95 in this cluster. It appears that over the initial 20 organisms, 18 are retrieved in the walktrap

cluster 0. Of these 18 organisms, 14 have their 3 key GHs gathered into a conserved locus. For the 4 organisms lacking at least one key GH, it appears that one have an incomplete context (i.e the locus is interrupt by the end of the genome file), an another have an assembly gap in its assemblage, the two other are false positive due to duplication non essential genes (i.e transporter) (Table 2).

Table 2: GHs found in the different potential PUL gathered into the walktrap cluster 0.

Organisms	class	GH31	GH35	GH95	GH74	Comments
Cellvibrio japonicus	Gammaproteobacteria	X	X	X		
Sphingomonas sp. Root241	Alphaproteobacteria	X	X			assembly gap into the locus
Massilia sp. Root1485	Betaproteobacteria	X	X	X	X	
Massilia sp. Root335	Betaproteobacteria	X	X	X	X	
Massilia timonae CCUG 45783	Betaproteobacteria	X	X	X	X	
Massilia timonae	Betaproteobacteria	X	X	X	X	
Massilia sp. JS1662	Betaproteobacteria	X	X	X	X	
Xanthomonas phaesoli pv. diffenbachiae	Gammaproteobacteria	X	X	X	X	
Pelomonas sp. Root1444	Betaproteobacteria		X	X	X	incomplete context due to the end of the genome file
Duganella sp. Leaf61	Betaproteobacteria	X	X	X		
Duganella phyllospharae	Betaproteobacteria	X	X	X		
Janthinobacterium sp. CG23_2	Betaproteobacteria	X	X	X		
Duganella sp. Leaf126	Betaproteobacteria	X	X	X		
Pelomonas sp. Root405	Betaproteobacteria		X	X		false positive due to the presence of many transporter (4 genes)
Pelomonas sp.	Betaproteobacteria		X			false positive due to

Root1217						the presence of many transporter (4 genes)
Teredinibacter turnerae T7901	Gammaproteobacteria	X	X	X		
Saccharophagus degradans 2-40	Gammaproteobacteria	X	X	X		
Microbulbifer thermotolerans	Gammaproteobacteria	X	X	X		

It should be noted that the Cellvibrio japonicus Xygul locis lacks a xyloglucanase activity (EC 3.2.1.151) which is essential for the degradation of the xyloglucan backbone (Larsbrink et al., 2014). However, this activity is found in the CAZY family GH74. In 7 organisms, the 3 key GHs are associated with a gene annotated as GH74 by DBCAN. Moreover, NetSyn unveils other genes that are involved in xylose metabolism or degradation generally. For instance, in the putative Xygul we have analysed, specific transporters (ABC sugar transporter, xylose ABC transporter) or enzymes (D-xylose-1-dehydrogenase) were found (Stephens et al., 2007). These findings emphasize that the locus we detected in the 14 organisms, gathered into NetSyn walktrap cluster 0, is dedicated to xyloglucan degradation. Organisms sharing a potential xyloglucan degradation locus belong to different classes as alphaproteobacteria, betaproteobacteria and gammaproteobacteria (Table 2). It shows that the conserved synteny in this Netsyn cluster is more due to a functional conservation rather than taxonomic conservation.

Automatic detection of the *Xygul* PUL in many different phyla was possible using NetSyn. Whereas in the litterature, this PUL is only described in β- and ɣ- proteobacteria, our analysis demonstrated that this PUL is also potentially present in alphaproteobacteria (Table 2). Here we have shown that NetSyn is able to group together enzymes acting in interaction into the same degrading system. Functional characterisation is needed to confirm our prediction.

Discussion & Conclusions

We have shown that NetSyn is able to divide more precisely homologous sequences classified into families than a method based on actif site modelling as ASMC. NetSyn is also able to group sequences disregarding their origin (i.e taxonomic rank of the organism). NetSyn is able to pinpoint pathways yet to be discovered. These putative pathways, never observed in any organisms, could be guessed by the annotation provided by NetSyn in the genomic context of the target genes. Moreover, NetSyn is able to highlight groups of non homologous proteins but belonging to a pathway not yet observed into some organisms. We have also demonstrated that NetSyn is able to group together enzymes acting in interaction into the same degrading system but without any homologous relationships and sequence similarity. Their grouping is due to their

genomic context conservation which reflects their functional conservation. But functional characterisation is needed to confirm our prediction.

NetSyn is a great tool to help functional assignment of protein sequences that are still of unknown function. Classical tools used sequence comparison to assign functions. However, even with recent improvement in homolog search accuracy (Buchfink et al., 2015), there are still 23 % of protein families with uncharacterized function in databases (Mistry et al., 2021). On the other hand, many known activities have not been associated with a protein sequence. In their review, Sorokina et al., show that more than 1000 EC numbers, called orphan reactions, are not linked to a specific gene (Sorokina et al., 2014). To link genes with orphans reactions, methods based on protein sequence similarity are not suitable, but methods able to group proteins according to their function without taking into account sequence similarity is required. NetSyn, which gathers sequences depending on their context conservation, could be the ideal tool for tackling this issue.

Synteny conservation is one promising method that can be used to challenge the issue of linking genes with reaction. Also network generating and clustering allows to extract interesting information and organise results in a meaningful way. Recent efforts based on synteny conservation were published, like EFI-GNT which analysed up to hundred of genomes and generated vast amounts of data (Zhao et al., 2016). Despite the accuracy of the method, proper data verification is needed. EFI-GNT uses genomic context to predict gene function but such analysis may be biased by the fact that their network is based on sequence similarity. Also the context exploration is made through the PFAM classification which does not cover the entire protein sequence space. NetSyn, on the contrary, disregards database classification, but uses MMseqs2 to create its own homologous families. To our knowledge, NetSyn is the only tool able to handle a set of proteins without any common evolutionary origin whereas the other methods start from an homologous family. NetSyn paves the way to new and innovatives tools to solve the challenging task of predicting new enzymatic activities.

The main limitation of NetSyn is the fact that it may be biased by a selection of sequences from organisms classified at the same low taxonomic range (i.e organisms from the same genus). Indeed the lack of genomic diversity in the input set might lead to detected a synteny only due to closely related organisms rather than functional conservation through evolution. To circumvent this issue, we proposed to reduce the redundancy by merging nodes belonging to the same NetSyn clusters which share the same criteria like a taxonomic rank or a given metadata.

The main guiding idea of NetSyn is that genes sharing a common genomic context are more likely to share the same function. But this comparison must be independent of their sequence similarities. Here, the relationship between two proteins in the network depends only

on the number of homologous genes into their corresponding genomic context. Therefore NetSyn is a suitable tool to explore either the functional diversity of enzyme families, either to highlight conserved loci containing genes of different evolutionary origins implied into the same pathway or in the same degradation/biosynthesis process. NetSyn is freely available at https://github.com/labgem/netsyn.

References

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

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

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176

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).

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

Ficko-Blean, E., Préchoux, A., Thomas, F., Rochat, T., Larocque, R., Zhu, Y., Stam, M., Génicot,

S., Jam, M., Calteau, A., Viart, B., Ropartz, D., Pérez-Pascual, D., Correc, G., Matard-Mann, M., Stubbs, K.A., Rogniaux, H., Jeudy, A., Barbeyron, T., Médigue, C., Czjzek, M., Vallenet, D., McBride, M.J., Duchaud, E., Michel, G., 2017. Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. Nat. Commun. 8, 1685. https://doi.org/10.1038/s41467-017-01832-6

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42, D222-230. https://doi.org/10.1093/nar/gkt1223

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

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

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

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457-462. https://doi.org/10.1093/nar/gkv1070

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., Apweiler, R., 2005. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 33, D29-33. https://doi.org/10.1093/nar/gki098

Karp, P.D., 2004. Call for an enzyme genomics initiative. Genome Biol. 5, 401. https://doi.org/10.1186/gb-2004-5-8-401

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

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
Mudgal, R., Sandhya, S., Chandra, N., Srinivasan, N., 2015. De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. Biol. Direct 10. https://doi.org/10.1186/s13062-015-0069-2
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
O Sheridan, P., Martin, J.C., Lawley, T.D., Browne, H.P., Harris, H.M.B., Bernalier-Donadille, A., Duncan, S.H., O'Toole, P.W., P Scott, K., J Flint, H., 2016. Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes. Microb. Genomics 2, e000043. https://doi.org/10.1099/mgen.0.000043
Orth, J.D., Palsson, B.Ø., 2010. Systematizing the generation of missing metabolic knowledge. Biotechnol. Bioeng. 107, 403–412. https://doi.org/10.1002/bit.22844
Osterman, A., Overbeek, R., 2003. Missing genes in metabolic pathways: a comparative genomics approach. Curr. Opin. Chem. Biol. 7, 238–251. https://doi.org/10.1016/s1367-5931(03)00027-9
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
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
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
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
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., 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., 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

Terrapon, N., Lombard, V., Gilbert, H.J., Henrissat, B., 2015. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. Bioinforma. Oxf. Engl. 31, 647–655. https://doi.org/10.1093/bioinformatics/btu716

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

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

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

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., Yin, Y., 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 46, W95–W101. https://doi.org/10.1093/nar/gky418

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