NetSyn: 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 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 huge number of sequences cannot be achieved without the help of bioinformatics tools and the development of new strategies. Standard methods for assigning a biological function to a gene are based on sequence similarity. However, another way is to mine databases to identify conserved gene clusters (i.e. syntenies). Indeed 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 sequences. Homologous protein families are then computed to measure GC conservation between each pair of input sequences using a GC conservation score. A network is 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 a step further by providing additional subfamilies to those already 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

find a locus composed of 3 non homologous glycosyl hydrolases, involved in the degradation of xyloglucan, in 162 genomes.

Discussion. NetSyn is able to cluster proteins according to their genomic context, making it possible to establish functional links between proteins without considering their sequence similarity alone. 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 for pinpointing annotation errors that have been propagated across databases, and for suggesting annotations on proteins currently annotated as "unknown". 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 in 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). Moreover, this issue is compounded by current automatic annotation procedures, which still have a high false-positive rate. This is particularly true for methods that rely solely on sequence similarity, leading to annotation error rates of up to 80% for certain protein families (Schnoes et al., 2009). Classical tools used sequence comparison to assign functions. However, even with recent improvement in homolog search accuracy, 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 overcome the limitations of these methods, innovative strategies have been developed: combination of genome analysis and metabolic information (Chen et al., 2011; Kharchenko et al., 2006; Smith et al., 2012; Yamanishi et al., 2007), 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), and more recently, machine learning techniques based on natural language processing (Ofer et al., 2021).

One way to address the challenge of annotating gene function is to compare the Genomic Context (GC) of genes 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). GC 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 GC method computes distances, in number of bases or genes, between two adjacent genes transcribed in the same direction (Overbeek et al., 1999). Small distances are considered to be a good indicator of co-regulated gene groups, as it is the case 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 proteome of several genomes: genes A and B are functionally related if orthologs of A and B have 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, i.e., the same metabolic pathway. The gene neighbour method (Bowers et al., 2004) calculates the distance between two genes in different genomes: genes A and B are functionally related if orthologs of A and B across many organisms are nearby on the chromosome (also called gene neighbours). To avoid taxonomy bias, the studied genomes must be carefully selected and not too closely related (*i.e.* 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 all 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 by a graph where nodes correspond to proteins and edges to interactions. Such a network highlights the interactions between different protein families. 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 as input the protein sequences of the target genes and the ones of their genomic contexts: similarities in the NCBI's reference genomes database are search for. Only blast hits separated by a user-defined maximum distance (in bp) and belonging to the same organism and scaffold are kept. Another example is given by GCsnap method (Pereira, 2021), which retrieves the genomic contexts by searching for the flanking genes of the input sequences; then it computes protein families of all the collected genes based on their sequence similarities. It is also important to mention the Enzyme Function Initiative (EFI) (Oberg et al., 2023; Zallot et al., 2019) which has led to the development of a method based on the Genomic Neighborhood Network (GNN). It computes a Sequences 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).

None of the above methods focus on developing a network that prioritizes the conservation of genomic context over sequence similarity to explore the functional diversity of protein families. Here, we present NetSyn (Network Synteny), a tool designed to link proteins in a network according to the conservation of their genomic context. NetSyn offers a number of features to explore the synteny network, such as sequence clustering, visualisation of the genomic context of each sequence, and mapping of data from various resources. Our method quantifies the conservation of the genomic context of proteins from different genomes. The major originality of NetSyn over existing methods is that sequence comparison is not primarily based on the similarity they share, but rather on the conservation of their genomic contexts. This important feature enables genomic comparisons of both homologous sequences, such as an entire family of proteins, and non-homologous ones from several protein families. NetSyn has thus been validated on two kinds of proteins dataset. On an initial set of sequences from the homologous family β-keto acid cleavage enzyme (BKACE) family (Bastard et al., 2014), NetSyn was able to create iso-functional sub-families and to identify putative new pathways.The second set of proteins was made of sequences from three non-homologous families involved in the same degradation pathway. NetSyn was able to retrieve this known pathway in different organisms, including procaryotes for which the presence of this degradation pathway was not suspected. Overall, the NetSyn principle is based on the idea that proteins sharing a similar genomic context are more likely to share similar functions or to be involved in the same biological function, i.e. a metabolic pathway. Consequently, our method can be easily extended to various sets of proteins.

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.

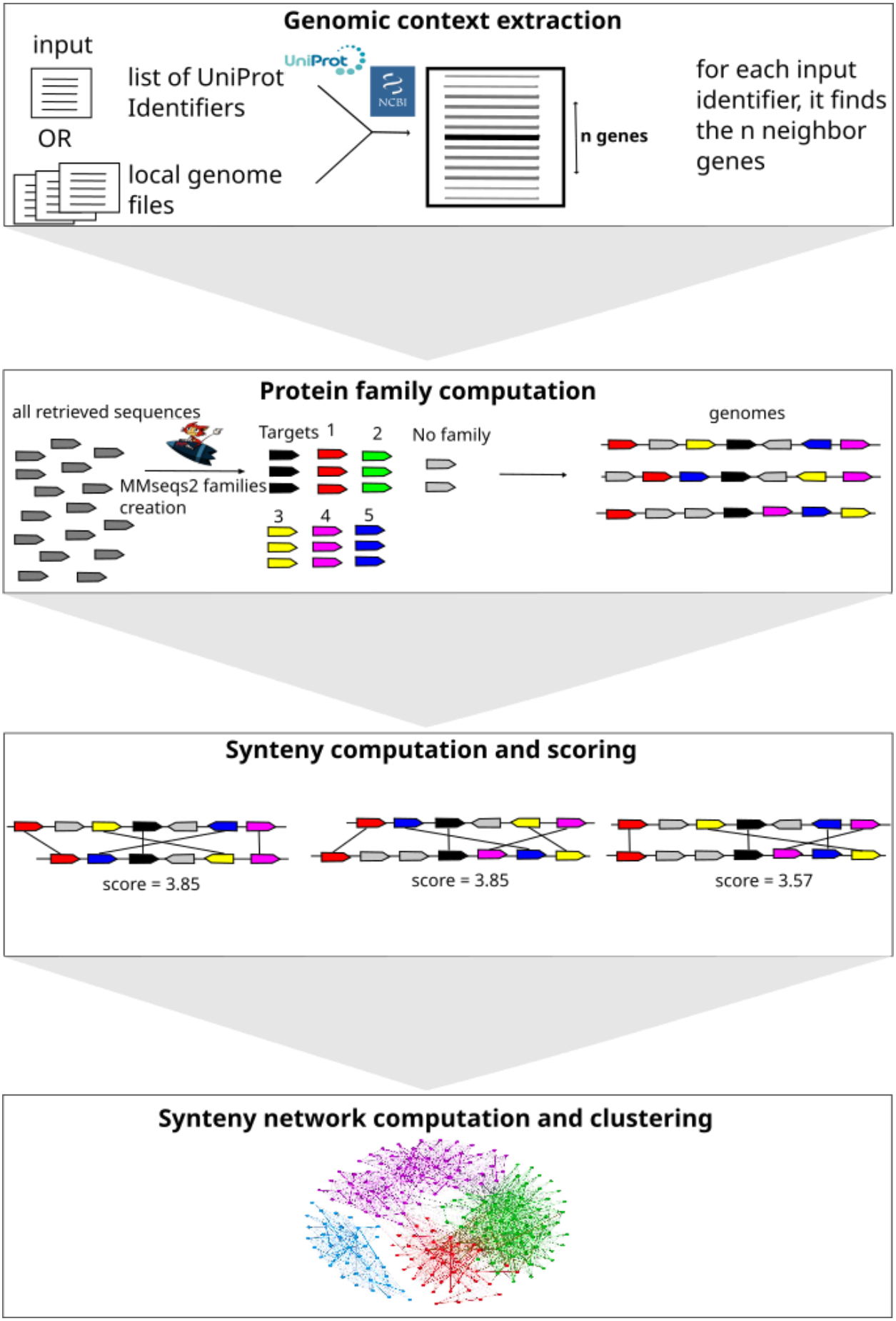
Genomic context extraction
input
list of UniProt Identifiers
OR
local genome files
UniProt
NCBI
n genes
for each input identifier, it finds the n neighbor genes
Protein family computation
all retrieved sequences
MMseqs2 families creation
Targets 1
2
3
4
5
No family
genomes
Synteny computation and scoring
score = 3.85
score = 3.85
score = 3.57
Synteny network computation and clustering

Figure 1 : *NetSyn Workflow*
The Uniprot accession of the target sequences (or the genome assembly files) are given to NetSyn as input. Then NetSyn gets the neighbour genes from the corresponding genome assembly files. The retrieved protein sequences are clustered in order to define protein families (default values: 30% of identity on 80% of coverage). All syntenies and scores are computed for each pair of target proteins. A graph is generated with target sequences as nodes. Edges are created if a synteny with a score above a threshold (3 by default) is found. At last, the network is divided using four different community detection algorithms (see text).

Genomic context extraction

NetSyn takes as input a list of UniProt accession numbers (hereafter called target proteins). The workflow uses the UniProt rest API to retrieve for each entry the corresponding nucleic accession number and download the genome assembly file (hereafter called the genome file) from the NCBI. NetSyn can also take as input local genome files listed in a tabulated file containing the following mandatory fields:
- protein_AC: protein target accession number in the genome file
- protein_AC_field: name of the field (qualifier) from which the protein accession number can be found in the genome file
- nucleic_AC: nucleic sequence accession number of the genome file
- nucleic_File_Format: format of the genome file (EMBL or Genbank format)
- nucleic_File_Path: path of the genome file stored locally.

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 also be used by Netsyn. These attributes will be displayed on the final synteny network (see below).

Genome files are used to extract the sequence of the target proteins and the Genomic Context (GC) of each target protein. The number of neighbouring genes depends on the window size parameter. The default value is set to 11, i.e. NetSyn takes into account 5 genes before and 5 genes after the target gene. NetSyn also recovers, from the genome files, the complete taxonomic lineage of the corresponding organisms. This taxonomic information can be subsequently used to remove the redundancy of the final graph (see section “Merging nodes and network reduction”).

Protein family computation

All the retrieved proteins (target proteins and proteins from the neighbour genes) are clustered using MMseqs2 (Steinegger and Söding, 2017), based on their sequence identity and alignment coverage percentages. Default values are set to 30% of identity on 80% of coverage (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

Computation of the genomic context conservation (*i.e*. synteny conservation) between two target proteins is made using an exact graph-theoretical approach (Boyer et al., 2005). From the two compared genomes containing the target proteins (T_A and T_B), two graphs (N_A and N_B) are extracted containing the target proteins and their neighbour genes as vertices ($G_{A1\text{-}n}$ and $G_{B1\text{-}n}$, according to window size parameter), with edges representing their colocalization on the sequence. To account for potential gaps (*i.e.* genes within the conserved synteny that lack homologs in the compared genome), 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). Next, edges are added between N_A and N_B to represent homologous relationships between genes that belong to the same protein family computed previously. Finally, Common Connected Components (CCC) between the two graphs are identified by constructing a multigraph in which vertices of NA and NB are merged based on homologous relationship edges. The detected CCCs correspond to conserved synteny groups. Figure 2 presents an example of the algorithm, showing the extracted genomic contexts around the T_A and T_B genes coding the target proteins (panel A) and the resulting detected synteny (panel B). G_{A1} and G_{B2} genes are not members of the conserved synteny as the number of genes between G_{A1} and T_A in the genome A is greater than the gap parameter threshold.

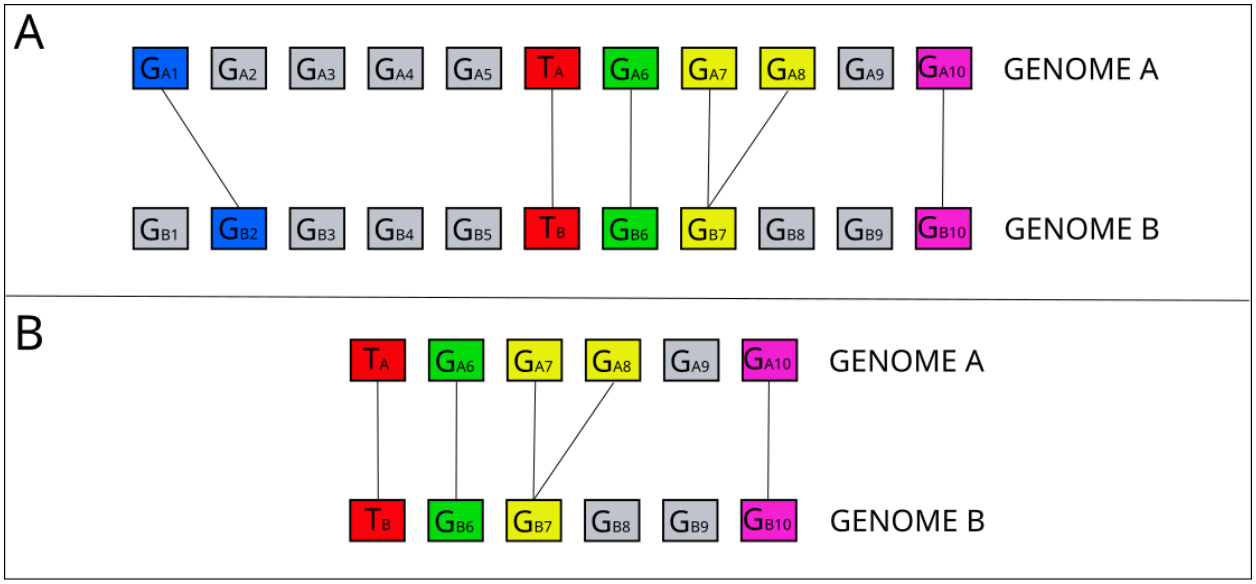

Figure 2: *Example of synteny computation between two target genes.*
A) Genomic context extraction from the two target genes: T_A from genome A and T_B from genome B. In this example the window size is set to 11. It means that NetSyn takes into account 5 genes upstream and 5 genes downstream the target genes to extract the genomic contexts in both genomes. Homologous genes between the two genomic contexts are represented by coloured rectangles and linked by an edge. B) Detected synteny. The synteny predicted by NetSyn contains 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 members of the conserved synteny as the number of genes between G_{A1} and T_A in the 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 among a total of 12 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 the total number of genes in the both genomes 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 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 the target proteins, NetSyn builds a synteny network whose nodes represent the target proteins. Two target proteins are linked by an edge if a conserved synteny is detected between them. The weight of the edge is equal to the synteny score, highlighting the most conserved synteny. To consider only significant syntenies, NetSyn, by default, retains those with a score of at least 3. This threshold is adjustable by the user. Target proteins that are not linked to any other target proteins will not be included in the final network.

In order to group target proteins sharing similar genomic contexts, several community detection algorithms have been 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 flows in graphs (it simulates random walks within a graph). The Walktrap algorithm (Pons and Latapy, 2005) is also based on a random walk and works well with large networks (> 1000 vertices). Louvain (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 4 algorithms are independently used to compute groups of target proteins sharing similar genomic contexts. On the synteny network, the corresponding clusters are shown using one specific colour for target proteins belonging to the same group. By default, results of the walktrap method are presented; the results of the three other clustering methods can also be visualised by changing the default parameters of the Netsyn graphical interfaces (see below).

NetSyn has been developed in Python 3 and can be installed on any Linux PC or server. The analysis of a set of around 700 sequences on a laptop with a single processor (intel core i7) is completed in about 300 seconds. The limiting step in terms of runtime is downloading the genome files.

Cluster quality

In the clustering framework based on the conservation of genomic contexts, a cluster is considered coherent when most of the nodes in the same cluster are interconnected. To achieve this goal, we use the alpha index which gives an idea of the connectivity within a specific cluster. The alpha index is the ratio between the number of loops (i.e., cycles which are paths in a cluster starting and ending at the same node and passing through nodes only once) and the estimated total number of loops in the network. This index is calculated using the following formula: alpha = (e-v)/((v(v-1)/2)-(v-1)) with e = number of edges, v = number of vertices. A value of 1 indicates a completely connected cluster while a value of 0 means a completely disconnected cluster.

NetSyn graphical interfaces

NetSyn network visualisation

The final NetSyn's output is a network. We provide two different file formats for exploring and analysing the network : a graphML file that can be opened with a network visualisation tool such as Gephi or Cytoscape, and an HTML file that can be opened in a Web browser (Fig. 3).

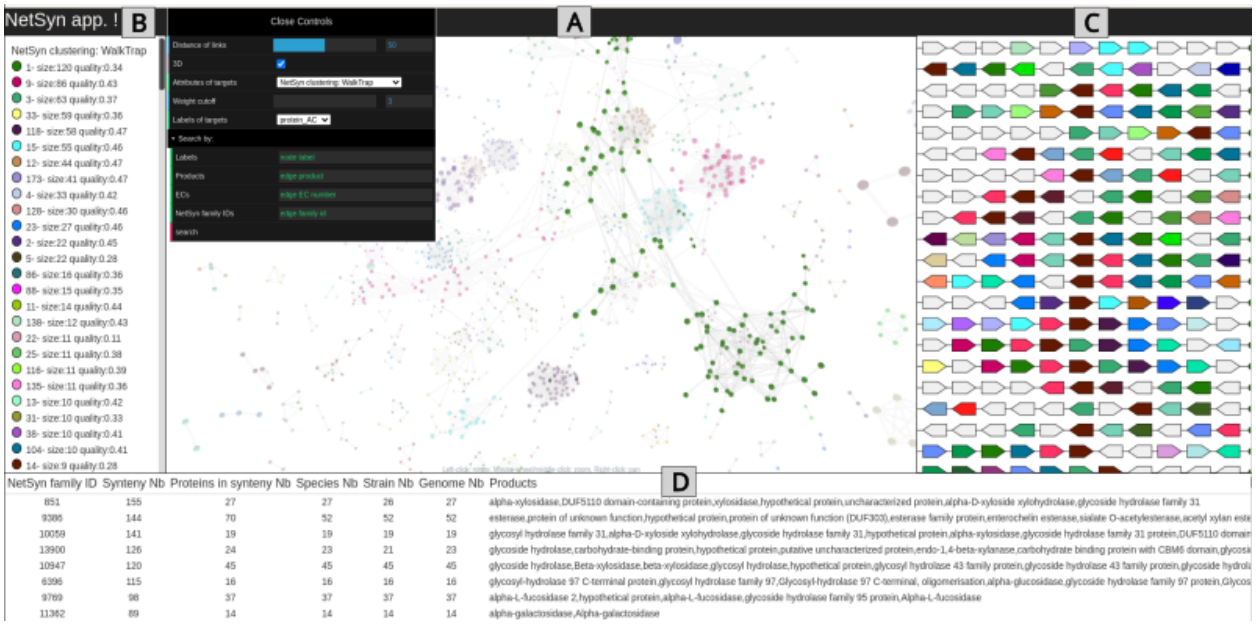

Figure 3: Visualisation of the NetSyn results.

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, taxonomy ranks or user-defined metadata. Nodes can be selected using the "Search by" functionality with the following search fields: Labels (names of the target genes), Products (functional descriptions), ECs (annotated EC numbers), NetSyn family IDs (identifiers of the protein families). In this panel, the network can be zoomed in or out, and rotated.
Panel B provides the color code legend for the currently selected 'Attributes of targets' attribute ordered by decreasing number of protein targets. The user can click on a specific coloured dot (here, cluster 1 obtained with the walktrap method) to highlight all the corresponding nodes in the network and display their genomic context in the right panel (C). Each arrow represents a gene and the same colour is used for homologous genes (i.e genes belonging to the same protein family). The different genomic contexts are centred on their respective target protein. When the user moves the mouse pointer over a gene, the following information is displayed in a pop-up window: the protein family identifier, strain name, UniProt accession number, locus tag, product, gene name and EC_number if any. In the bottom panel (D), a summary table for the different protein families found into a cluster is shown with the following columns: NetSyn family ID (protein family identifier), Synteny Nb (the number of times this family is found in a synteny), Proteins in Synteny Nb (number of corresponding proteins of the family), 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 its number of occurrences.

The HTML view is divided into 4 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 threshold (3 by default) has been identified between the two target proteins. Nodes can be colored according to different attributes, such as Walktrap cluster identifiers by default, as well as genome taxonomy or user-defined metadata. The control panel at the top of the A panel allows the user to select different attributes and information to be displayed on the synteny network (see Fig. 3 captions). The left panel (B) shows the legend of the network which depends on the parameters selected by the user in the control panel. In the case of the Walktrap algorithm, the identifier, the number of protein targets and the alpha index (see below) of each cluster are shown (Fig. 3B). The selection of one coloured spot of the B panel, automatically highlights, on the synteny network (panel A), nodes having the corresponding attributes. In addition, the genomic contexts of the corresponding target proteins are displayed on the right panel (C). Each context is centred on the target protein gene, surrounded by other genes from the conserved synteny. In this view, homologous genes are of the same colour (*i.e.*, they are members of the same protein family). The lower panel (D) gives, for the selected group of target proteins (panel B), a summary table of the different protein families found among the conserved syntenies (number of proteins, number of corresponding species, strains and genomes, and functional annotations of the proteins in the family).

Merging nodes and network reduction

In some cases, the final network can be huge and difficult to analyse. To get around this problem, we have added an option to reduce the network. Nodes identified as belonging to the same synteny cluster using only one of the four available clustering methods and sharing a given

property, will be merged into a unique node. The property can be a taxonomic rank or a family classification (Fig. S1). The merging is applied on one property at a time in a single NetSyn run. The “merged nodes” are the result of the union of one NetSyn cluster obtained with the chosen clustering method and one property.

Results & Discussion

The NetSyn method has been tested on two datasets. The first one is a family of homologous proteins, the β-keto acid cleavage enzymes family, that has been well studied with experimental validations. The other dataset contains a mixture of three evolutionarily distinct families that are known to be co-localized and interact to degrade xyloglucan.

Dividing an homologous family

The β-keto acid cleavage enzymes 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). The profiles of activity issued from the enzymatic screening of 124 BKACE representative sequences on 20 substrates showed a high correlation with the profiles of active sites. Indeed, 725 proteins of the BKACE family were divided into 7 main groups based on a combination of different clustering methods. One of them was the structural classification of their active sites computed using the Active Sites Modeling and Clustering (ASMC) method (de Melo-Minardi et al., 2010). The ASMC groups correlate well with the biochemical characterizations. In this analysis, we therefore wanted to compare the ASMC and NetSyn clusters to check whether there was a good overlap between the two methods.

We submitted the same set of 725 UniProt entries to NetSyn, 159 of which could not be associated with an ENA file because these UniProt entries were obsolete. Among the remaining 566 entries, 86 had no relevant conserved genomic context (*i.e.* with a synteny score >=3). Therefore, 480 input entries were retrieved into the final network. The Louvain algorithm (see Mat. & Meth. section) leads to a minimum number of clusters (33) whose size is between 2 to 68 sequences (Table 1). Most of these clusters correlate well with the 7 clusters already defined by the ASMC method which have been experimentally validated.

ASMC / NetSyn	G1	G2	G3	G4	G5	G6	G7	Total	max coverage %
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	

Table 1: Comparison between ASMC groups and Louvain clusters

	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 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 process non homologous but functionally related genes. Indeed, half of the genes in prokaryotes and a minor fraction of the genes in eukaryotes are physically linked on chromosomes to form clusters of genes involved in a specific function (Nützmann et al., 2018). The most widely studied genetic organisation is the operon found in Prokaryotes (Jacob and Monod, 1961) in which genes are transcribed in a single polycistronic mRNA. Operons are often conserved across species by vertical inheritance, and it is therefore quite easy to computationally predict them.

Polysaccharides Utilisation Locis ((PULs) have been described in bacteroidetes and contain physically linked genes, involved in the carbohydrate metabolism and two proteins composing a transport system: a TonB dependent receptor (namely susC) and a carbohydrate binding protein (namely susD). Therefore the strategies for PULs prediction are based on the detection of the susC/susD pair associated with a cluster of Carbohydrate Active Enzymes (CAZymes) (Terrapon et al 2015). However the SusC/SusD paradigm has been called in question by the discovery of other types of sugar transport systems associated with clusters of CAZyme (Larsbrink et al 2014; O Sheridan et al, 2016) and by the fact that the SusC/SusD system is not necessarily localised near the CAZymes genes (Ficko-Blean et al 2017). Therefore, to detect gene clusters or PULs, methods such as NetSyn, which highlights similar gene context (GC) containing non-homologous genes, is required.

Our goal here was to focus on the interconnectivity between the GH proteins from different CAZy families involved in xyloglucan degradation. Xyloglucans (XyG) constitute a family of ubiquitous and abundant plant cell wall polysaccharides. They are composed of β(1->4)-glucan backbone appended with branching α(1->6)-xylosyl moieties which can be extended with additional glycosides (Fig 5 panel B) (Grondin et al 2017). Xyloglucan can be degraded by a complex gene locus in the saprophyte Gram-negative *Cellvibrio japonicus* called a XyGul (Xyloglucan Utilisation Loci). This locus is composed of 3 glycosyl hydrolases (GH) of the GH31, GH35 and GH95 families defined in the CAZy databases (Drula et al., 2022) and a TonB-dependent receptor (TBDR) (Larsbrink et al 2014). This particular locus has also been found in several strains of Xanthomonas (Vieira et al., 2021).

We used the three key gene (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). A total of 15452 protein candidates (2212 homologous to xyl31A, 1756 homologous to bgl35A and 11484 homologous to afc95A) were submitted to NetSyn. Of these 15452 entries, 251 could not be associated with a GenBank identifier and 379 with a genome flat file. Moreover, NetSyn did not find conserved genomic context 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 a number of sequences ranging from 910 to 3005, 7 clusters with a number of sequences ranging from 106 to 509, 15 clusters with a number ranging from 20 to 89 and 282 clusters with a number of sequences ranging from 2 to 18 (figure 4A).

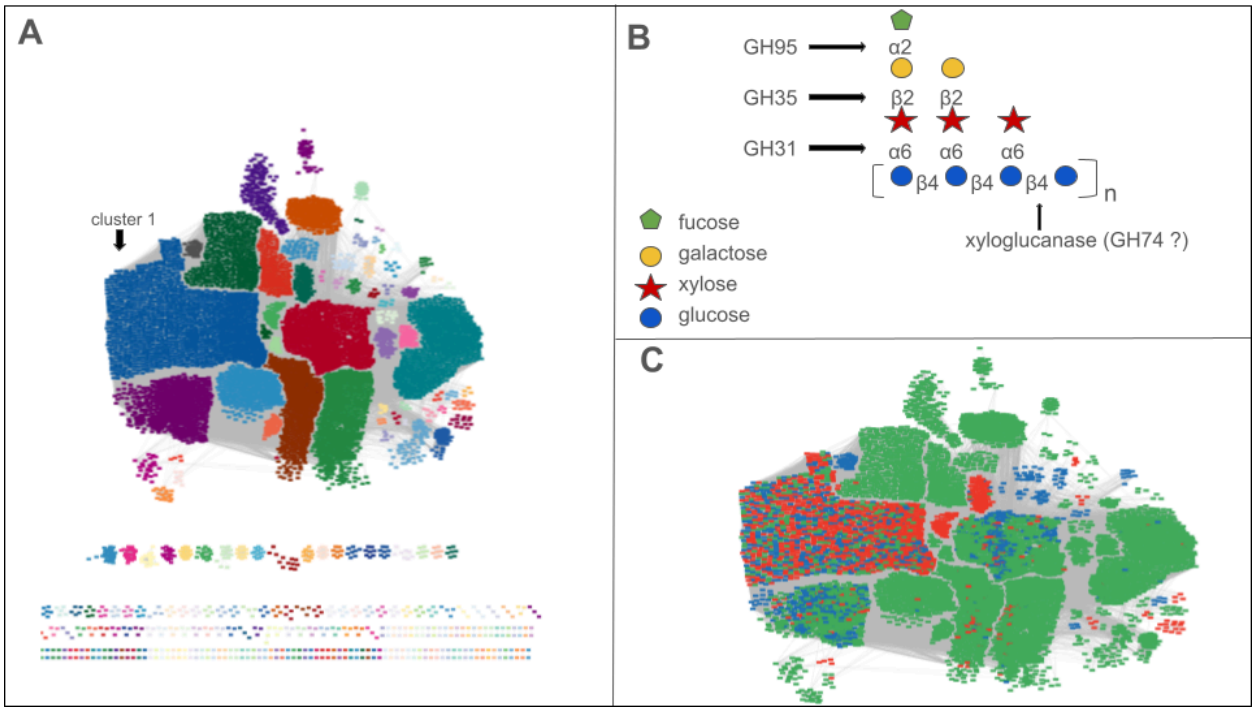

Figure 4:

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 together sequences belonging to the 3 GH families (clusters 1, 7, 6, 5, 2 and 10), which are most often localized in different locus. Indeed, only cluster 1 brings together the 3 different GHs (GH31, GH35 and GH95) from colocalized genes including the 3 initial GH from *C. japonicus*. By investigating the different loci of cluster 1, we were able to identify 166 potential XyGuls in 162 distinct genomes belonging to different classes of Proteobacteria (alpha, beta and gamma). Forty Xygul loci have been described in various Xanthomonas strains (Vieira et al., 2021). Of these 40 loci, 19 were found by NetSyn. The others were missed because the proteins involved in the Xygul locus were either not present in Uniprot, or had less than 30% identity with the sequences of *C. japonicus* (therefore they were not included in the initial dataset).

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

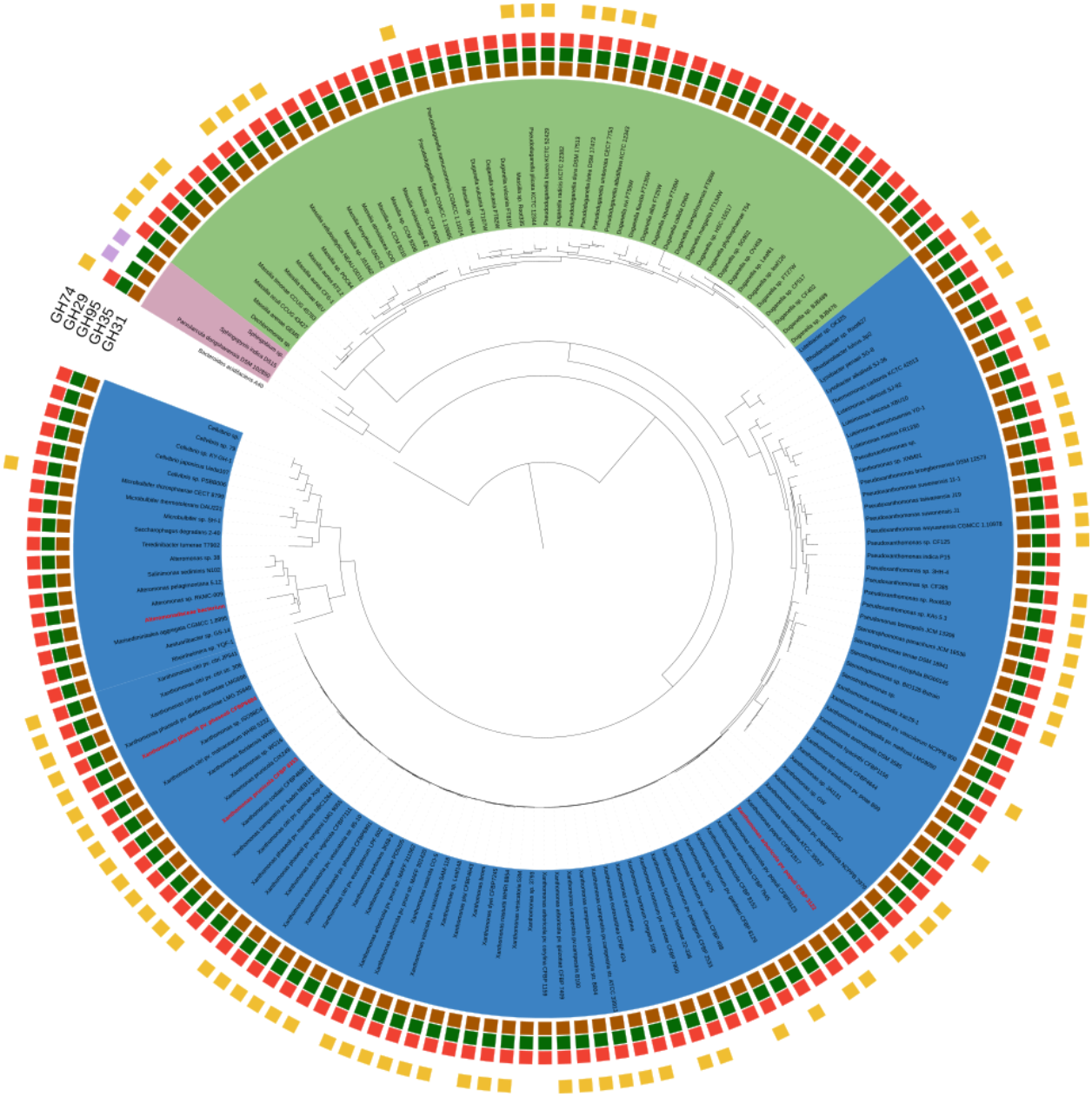

Figure 5:

The tree was created from the 16S sequences of organisms in which we detected a Xygul locus similar to that of C. japonicus. A 16S sequence from Bacteroides acidifaciens A40 was used as an outgroup. Starting from the center to the outside of the circle, the first layer shows the name of the organisms: the same genus is of one specific background colour (proteobacteria in pink, betaproteobacteria in green and gammaproteobacteria in blue). The second layer shows organisms containing GH31 (brown squares), the third one organisms containing GH35 (green squares), the fourth one organisms containing GH95 (red squares), the fifth one organisms containing GH29 (purple squares) and the last layer, organisms containing GH74 (yellow squares).

Moreover, NetSyn unveils other genes that are involved in xylose metabolism (or more generally in degradation), i.e., xylulokinase and acetylesterase. For instance, specific transporters (ABC sugar transporter, ABC xylose transporter) or enzymes (D-xylose-1-dehydrogenase) were found in the putative Xygul loci previously identified (Stephens et al., 2007). These results show that the locus detected in the 160 organisms gathered in NetSyn cluster 1, is dedicated to the degradation of xyloglucan. Indeed, organisms for which a xyloglucan degradation locus has been predicted by Netsyn belong to different classes: alphaproteobacteria, betaproteobacteria and gammaproteobacteria (Fig 5). This clearly shows that the synteny conserved in this Netsyn cluster is due more to a functional conservation than to taxonomic conservation.

The use of Netsyn allowed the detection of the *Xygul* locus in many different phyla, including alphaproteobacteria (Table 2). Finally, we have shown here that NetSyn is able to group together enzymes acting in interaction into the same degrading system. However, functional characterisation is still required to confirm our predictions.

Conclusion

We have presented NetSyn, a new tool for linking sequences in a network according to the conservation of their genomic context, and we have shown its application on 2 specific protein datasets.. NetSyn fills the gap left by current methods based solely on sequence similarity or active site modelling, such as ASMC, for calculating groups of isofunctional proteins regardless of the taxonomic rank of the organism to which they belong.

By also grouping sequences according to the conservation of their genomic context, NetSyn seems to be an ideal tool to help with the functional assignment of protein sequences whose function is still unknown. Results obtained by applying NetSyn to two protein datasets show that our tool is capable of gathering enzymes acting in the same degrading system, but without any sequence similarity. Their grouping is in fact due the conservation of their genomic context which reflects their functional conservation. The mixing of homologous and non-homologous protein sequences as input of methods such as NetSyn is necessary when one wants to study genomic structures such as operons, in which several genes involved in a specific cellular function (*i.e.* metabolic pathway, regulation, etc.) are co-localized in the same locus. NetSyn is therefore an effective tool for highlighting groups of non homologous proteins belonging to a biological system that has not yet been observed in certain organisms, for example, a metabolic pathway that has yet to be discovered. We have shown that these new putative enzymatic activities could be guessed by the annotation provided by NetSyn in the genomic context of the target genes. As previously stated, recent efforts based on synteny conservation have been published, such as the EFI-GNT initiative, which makes it possible to analyse up to hundreds of genomes and generate large quantities of data (Zhao et al., 2016).

Despite the accuracy of this method, the analysis is directed by the construction of a Sequence Similarity Network (SSN) based only on sequence homology rather on genomic context conservation of the studied families. As a result, EFI does not allow an analysis combining proteins of different evolutionary origins, whereas NetSyn can highlight conserved contexts between sequences with different evolutionary origins. Furthermore EFI's clustering depends on a threshold provided by the user and therefore requires a good knowledge of the family being studied, whereas NetSyn creates its clusters without any preconceived ideas.

The 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 could lead to the detection of synteny due solely to closely related organisms rather than functional conservation through evolution. To overcome this problem, we proposed to reduce the redundancy by merging nodes belonging to NetSyn clusters sharing the same taxonomic rank or given metadata. Another current limitation is the number of sequences that NetSyn can handle: above 10,000 sequences, the graphML files may be too large for Gephi or Cytoscape to open, and the html output too large for a browser to open.

To our knowledge, NetSyn is the only tool currently capable of calculating a network in which the relationship between two proteins depends only on the number of homologous genes into their corresponding genomic context. Our tool can therefore be used to explore either the functional diversity of enzyme families, or to identify conserved loci containing genes of different evolutionary origins involved in the same biological pathway (i.e., the same degradation/ biosynthesis process). NetSyn paves the way to new and innovative tools to solve the challenging task of predicting new enzymatic activities.

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

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

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

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

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

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

694 databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput.
695 Biol. 5, e1000605. https://doi.org/10.1371/journal.pcbi.1000605
696 Smith, A.A.T., Belda, E., Viari, A., Medigue, C., Vallenet, D., 2012. The CanOE strategy:
697 integrating genomic and metabolic contexts across multiple prokaryote genomes to find
698 candidate genes for orphan enzymes. PLoS Comput. Biol. 8, e1002540.
699 https://doi.org/10.1371/journal.pcbi.1002540
700 Sorokina, M., Stam, M., Médigue, C., Lespinet, O., Vallenet, D., 2014. Profiling the orphan
701 enzymes. Biol. Direct 9, 10. https://doi.org/10.1186/1745-6150-9-10
702 Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for
703 the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028.
704 https://doi.org/10.1038/nbt.3988
705 Stephens, C., Christen, B., Fuchs, T., Sundaram, V., Watanabe, K., Jenal, U., 2007. Genetic
706 Analysis of a Novel Pathway for d-Xylose Metabolism in Caulobacter crescentus. J.
707 Bacteriol. 189, 2181–2185. https://doi.org/10.1128/JB.01438-06
708 Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L.,
709 Fang, T., Doncheva, N.T., Pyysalo, S., Bork, P., Jensen, L.J., von Mering, C., 2023. The
710 STRING database in 2023: protein-protein association networks and functional
711 enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 51,
712 D638–D646. https://doi.org/10.1093/nar/gkac1000
713 Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A.,
714 Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C., 2017. The STRING
715 database in 2017: quality-controlled protein-protein association networks, made broadly
716 accessible. Nucleic Acids Res. 45, D362–D368. https://doi.org/10.1093/nar/gkw937
717 Tamames, J., 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 2,
718 RESEARCH0020. https://doi.org/10.1186/gb-2001-2-6-research0020
719 Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G.,
720 Scarpelli, C., Médigue, C., 2006. MaGe: a microbial genome annotation system
721 supported by synteny results. Nucleic Acids Res. 34, 53–65.
722 https://doi.org/10.1093/nar/gkj406
723 van den Belt, M., Gilchrist, C., Booth, T.J., Chooi, Y.-H., Medema, M.H., Alanjary, M., 2023.
724 CAGECAT: The CompArative GEne Cluster Analysis Toolbox for rapid search and
725 visualisation of homologous gene clusters. BMC Bioinformatics 24, 181.
726 https://doi.org/10.1186/s12859-023-05311-2
727 Vieira, P.S., Bonfim, I.M., Araujo, E.A., Melo, R.R., Lima, A.R., Fessel, M.R., Paixão, D.A.A.,
728 Persinoti, G.F., Rocco, S.A., Lima, T.B., Pirolla, R.A.S., Morais, M.A.B., Correa, J.B.L.,
729 Zanphorlin, L.M., Diogo, J.A., Lima, E.A., Grandis, A., Buckeridge, M.S., Gozzo, F.C.,
730 Benedetti, C.E., Polikarpov, I., Giuseppe, P.O., Murakami, M.T., 2021. Xyloglucan
731 processing machinery in Xanthomonas pathogens and its role in the transcriptional
732 activation of virulence factors. Nat. Commun. 12, 4049.
733 https://doi.org/10.1038/s41467-021-24277-4
734 Wu, H., Owen, C.D., Juge, N., 2023. Structure and function of microbial α-l-fucosidases: a mini
735 review. Essays Biochem. 67, 397–412. https://doi.org/10.1042/EBC20220158
736 Yamanishi, Y., Mihara, H., Osaki, M., Muramatsu, H., Esaki, N., Sato, T., Hizukuri, Y., Goto, S.,
737 Kanehisa, M., 2007. Prediction of missing enzyme genes in a bacterial metabolic
738 network. Reconstruction of the lysine-degradation pathway of Pseudomonas aeruginosa.
739 FEBS J. 274, 2262–2273. https://doi.org/10.1111/j.1742-4658.2007.05763.x
740 Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., Gupta, D., Gupta, V.,
741 Haseeb, M., Ihsan, M., Ivanov, E., Jayathilaka, S., Kadhirvelu, V.B., Kumar, M., Lathi, A.,
742 Leinonen, R., McKinnon, J., Meszaros, L., O'Cathail, C., Ouma, D., Paupério, J., Pesant,
743 S., Rahman, N., Rinck, G., Selvakumar, S., Suman, S., Sunthornyotin, Y., Ventouratou,
744 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