

RESEARCH ARTICLE

Open Access



# Robust structure measures of metabolic networks that predict prokaryotic optimal growth temperature

Adèle Weber Zendrera<sup>\*</sup>, Nataliya Sokolovska and Hédi A. Soula

## Abstract

**Background:** Metabolic networks reflect the relationships between metabolites (biomolecules) and the enzymes (proteins), and are of particular interest since they describe all chemical reactions of an organism. The metabolic networks are constructed from the genome sequence of an organism, and the graphs can be used to study fluxes through the reactions, or to relate the graph structure to environmental characteristics and phenotypes. About ten years ago, Takemoto et al. (2007) stated that the structure of prokaryotic metabolic networks represented as undirected graphs, is correlated to their living environment. Although metabolic networks are naturally directed graphs, they are still usually analysed as undirected graphs.

**Results:** We implemented a pipeline to reconstruct metabolic networks from genome data and confirmed some of the results of Takemoto et al. (2007) with today data using up-to-date databases. However, Takemoto et al. (2007) used only a fraction of all available enzymes from the genome and taking into account all the enzymes we fail to reproduce the main results. Therefore, we introduce three robust measures on directed representations of graphs, which lead to similar results regardless of the method of network reconstruction. We show that the size of the largest strongly connected component, the flow hierarchy and the Laplacian spectrum are strongly correlated to the environmental conditions.

**Conclusions:** We found a significant negative correlation between the size of the largest strongly connected component (a cycle) and the optimal growth temperature of the considered prokaryotes. This relationship holds true for the spectrum, high temperature being associated with lower eigenvalues. The hierarchy flow shows a negative correlation with optimal growth temperature. This suggests that the dynamical properties of the network are dependant on environmental factors.

**Keywords:** Metabolic networks reconstruction, Directed graphs, Graph topology, Correlation with environment

## Background

All living organisms rely on chemical reactions to exist, and the set of these life-sustaining chemical transformations is defined as metabolism. Because these reactions are mostly catalysed—accelerated—by enzymes, the transformation of organic molecules (substrates) into other chemicals (products) can directly be mapped by the enzyme set.

The development of metabolic databases such as KEGG [1] linking enzymes to their reaction pair — substrates/products— allows us to explore the structure of metabolism in general, and to investigate the structures of the metabolic graphs of particular organisms [2].

Flow is an inherent concept of metabolic reactions, going from substrates to products which then become substrates for other reactions. Directed graphs are therefore a natural way to model enzymes and chemical reactions [3]. The metabolic network of an organism is defined as the whole set of metabolic pathways. Since such a network is a (directed) graph, the elements of graph the-

\*Correspondence: [adela@nicoweb.com](mailto:adela@nicoweb.com)  
Sorbonne University, INSERM, NutriOmics F75013, France., 91, blvd. de l'Hôpital, 75013 Paris, France



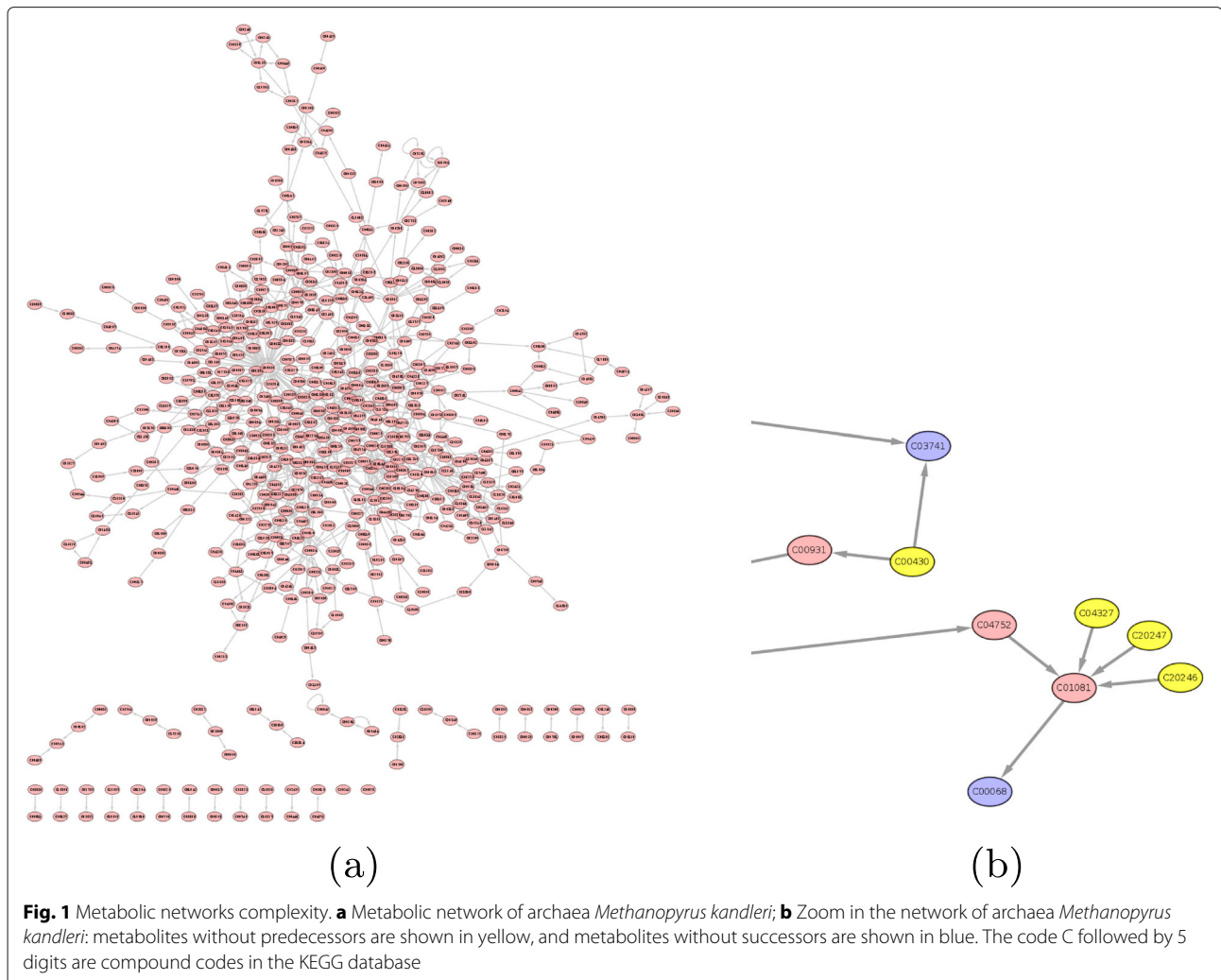
ory can be applied to study its properties. Understanding network topologies and their physical, chemical, and biological constraints is critical to decipher the function and evolution of cellular networks [4].

We focus on a metabolite-centered representation where the nodes of a graph are metabolites, and they are connected if an enzymatic reaction converting one metabolite into another exists. It is then simple to create metabolic networks that describe all chemical reactions of one or multiple organism(s) as a graph. Figure 1 provides an example of the topology of a directed graph for a bacterium. In a directed graph, some nodes are end-points (shown in blue), and some nodes are starting-points (in yellow). To keep our flow analogy in play, starting points compounds will be an input (e.g. from the medium) whereas end points are 'final' products of the complete pathways.

Recent studies take two main directions in the analysis of metabolic networks. The first one heavily relies on graph structural measures such as degree distribution,

clustering coefficient, path length, and centrality [5–7]. The second venue of research is based on the analysis of biomass dynamics inherent to metabolic networks, by trying to predict steady-state flux distributions. Various constraint-based techniques exist to solve this problem, e.g., flux balance techniques (FBA) [8].

It was reported by several studies (e.g., [9–12]) that the compounds of metabolic networks, the flow of substrates and products, and the overall pathway organisation are correlated to environmental variables and to phenotypical traits. The aim of these studies was to discover similarities and differences in the structural and functional properties of various organisms. It was noticed [9] that evolutionary changes in metabolic networks are mostly due to adaptation to changing conditions. So, Takemoto et al. [11] made an attempt to explore correlations between several structural properties of metabolic networks (such as edge density, power law degree exponent, clustering coefficient, and subgraph concentration) for 113 prokaryotes to their optimal growth temperature.



Ideally, metabolic networks require complex representations such as hypergraphs, since reactions in metabolic networks convert multiple reaction inputs into multiple outputs using other components [13, 14]. However, a reduced representation and algorithms on graphs can facilitate the analysis by addressing the fundamental biological concepts.

A number of graph theory approaches were proposed to study relations between the structure of metabolic networks and the environment. So, Borenstein et al. [10] stated that species whose environment are highly-predictable tend to have smaller sets of compounds that are exogenously acquired than those who live in variable conditions.

The network topology determines network functions [15], and the topology of a metabolic network is important in predicting the viability of mutant strains.

Metabolic networks are known to be extremely heterogeneous, and two networks of two different organisms are quite different [16]. At the same time, the metabolic networks were shown to be robust in the sense that elimination of several central nodes does not modify the functions of the networks [16]. A graph-based method to identify all minimal reaction sets in a metabolic network was considered in [17].

Our main motivation is to explore the structure of directed metabolic graphs of bacteria, and to relate it to phenotypes. In our experiments, we consider prokaryotic optimal growth temperature as a phenotype. In this article, our contribution is:

- We reconstruct metabolic networks for species considered by [11] in addition to several species that will increase the number of species in the growth temperature classes which were represented by too few species.
- We build and explore undirected and directed metabolic graphs; including all KEGG enzymes for the species or only those found in the so-called KEGG pathways.
- We propose to apply robust measures on directed complex graphs, namely largest strongly connected component, flux hierarchy and Laplacian spectrum, and we relate these measures to the environmental conditions.
- In our experiments, we have confirmed the results of [11], and we discuss the newly introduced metrics.

## Results

### Confirmation of the state-of-the art results of Takemoto et al. (2007)

In the study conducted by [11], undirected substrate graphs were constructed using KEGG metabolic pathways for 113 prokaryotes from four different growth

temperature classes (hyperthermophiles, thermophiles, mesophiles, and psychrophiles). They considered several properties of undirected graphs, and analysed the correlation between the graph properties and the optimal growth temperatures of the organisms. We implemented and tested three measures from their study: edge density, maximum likelihood estimate of degree exponent, and average clustering coefficient.

Since more than ten years have gone by, the databases have evolved significantly. We focused on directed graphs, and we reconstructed the metabolic networks without using pathways directly (see Methods). We obtained graphs for 100 out of the 113 species, filtering out nodes in such a way that we had metabolic reactions from known pathways only. We also added 128 additional species.

There are two main differences between our experiments and the ones from [11]. We use more species in our experiments, and we consider directed graphs. We still confirm most of the results from [11].

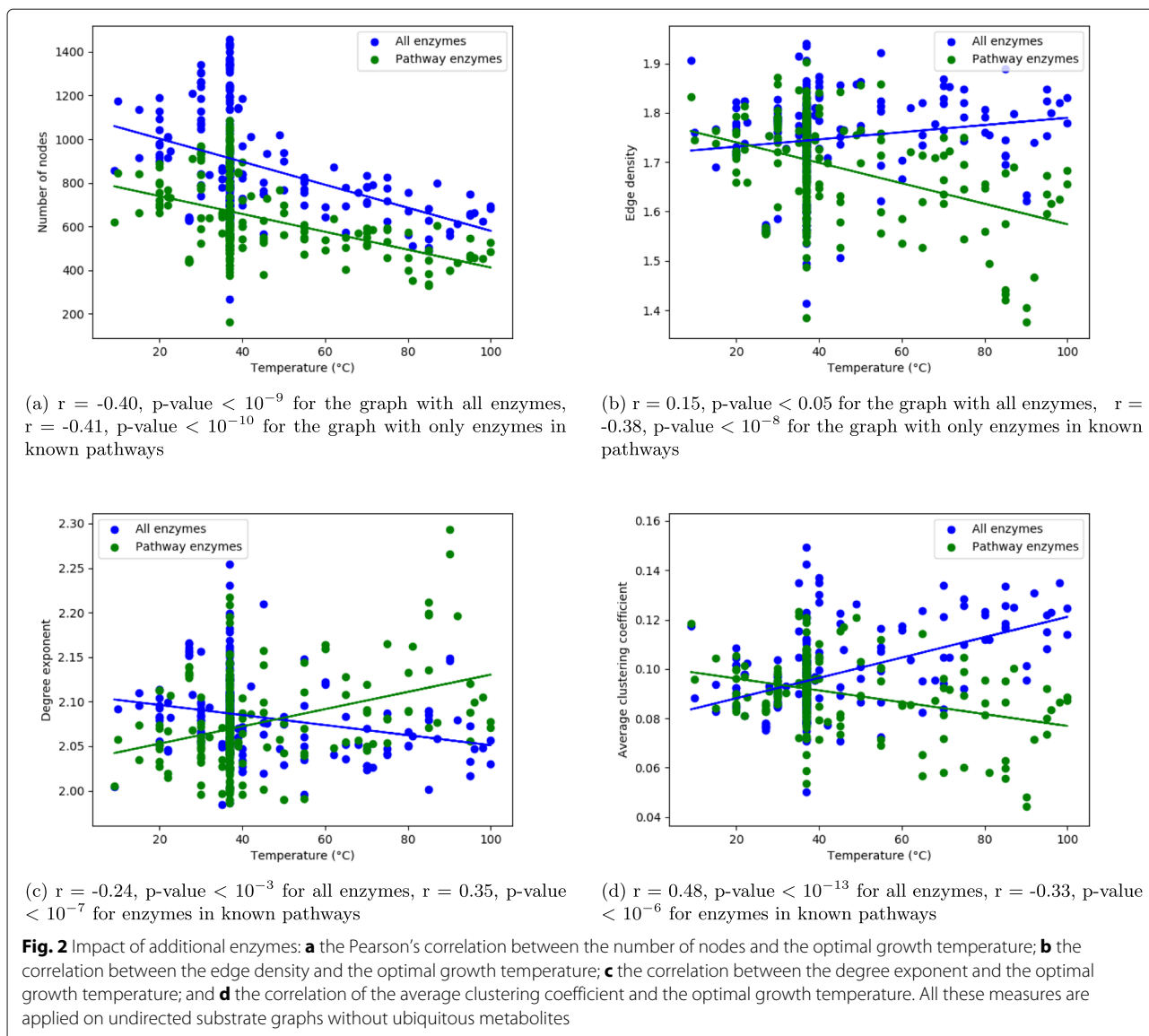
We apply a linear regression to estimate relation between the number of nodes in a metabolic network and the optimal growth temperature and deduce Pearson's correlation. We consider a correlation between the number of nodes and the optimal growth temperature, a correlation between edge density and optimal growth temperature, a correlation between degree exponent and optimal growth temperature, and a correlation of average clustering coefficient and optimal growth temperature. All these measures are applied to undirected substrate graphs without ubiquitous metabolites, to be as close to [11] as possible. As we can see from Fig. 2 (green lines), there is a significant negative correlation between the number of nodes and the temperature shown on Fig. 2a, and a negative correlation between edge density and the temperature illustrated on Fig. 2b. As shown on Fig. 2c and d, we find a significant positive correlation for the degree exponent estimate, and significant negative correlation for the average clustering coefficient with the optimal growth temperature of the species.

In general, we discovered the same tendencies as [11].

### Undirected graphs reconstructed from all enzymes and the impacts on graph properties

Here we compare the influence of taking all KEGG enzymes for a species with enzymes in known KEGG pathways only. Our method to build the metabolic graphs is different from [11] in that we consider all reactions that can be deduced from the species genes, and not only enzymes involved in known pathways. We have, therefore, found additional enzymes for all species.

On Fig. 2a, we notice a strong bias related to the number of nodes in respect to the growth temperature. Hence, ideally, graph properties are to be normalised by the number of nodes.



We wish to analyse the impact of the additional reactions. We observe a complete loss of the tendencies, what is shown in blue on Fig. 2b, c and d. All the correlations are inverted, albeit with lower correlations except for the average clustering coefficient, which has a stronger correlation than what [11] found.

We found a significant positive correlation between the proportion of enzymes that are not in a pathway and optimal growth temperature (Pearson's  $r = 0.26$ , data not shown), meaning that more new enzymes—edges—are added for thermophiles than non-thermophiles, adding more edges and therefore likely causing the correlation inversions.

When removing up to 40% of random nodes in the graphs with all enzymes, the trends stay significantly

correlated. This means that the pathway enzymes are specific and greatly modify the graph structure.

These differences could be due to a bias in KEGG pathways for hyperthermophiles, that could have less annotated and curated pathways than its more well-studied counterparts, and thus more enzymes not associated to pathways. Another hypothesis is that this difference could be explained by noise, since the number of nodes for hyperthermophiles is the smallest.

#### Robust directed measures to analyse cycles of metabolic networks

We focus on directed metabolic networks, and we are interested in finding relevant measures on directed graphs that can explain correlation with environmental variables

such as optimal growth temperature. We propose robust measures to analyse substrate graphs.

We considered two measures to study cycles in networks, the size of the largest strongly connected component which corresponds to the biggest cycle in a graph, and the flow hierarchy (see Methods). The flow hierarchy is defined as the number of nodes in a component that is not a part of the largest strongly connected component. So, the two explored measures are closely related.

As shown in Fig. 3, we found a significant negative Pearson's correlation between the size of the largest strongly connected component, normalised by the number of nodes, and the optimal growth temperature. We have also observed a significant positive Pearson's correlation between the node normalised flow hierarchy and the optimal growth temperature. These tendencies have been found for substrate graphs built with all enzymes and for substrate graphs with enzymes in known pathways: these measures are consistent in both cases, with similar correlations thus becoming robust measures to analyse correlation between metabolic network structure and environmental conditions. We consider that these measures are potentially more relevant to describe metabolic networks as they reflect directed graphs properties.

### Metabolic network Laplacian eigenvalues

We tested another directed network structural property to study connectivity. A temperature class network yields different graph structural properties related to connectivity, but the underlying description of these graphs is a compound flow. These properties are associated to the 'speed'

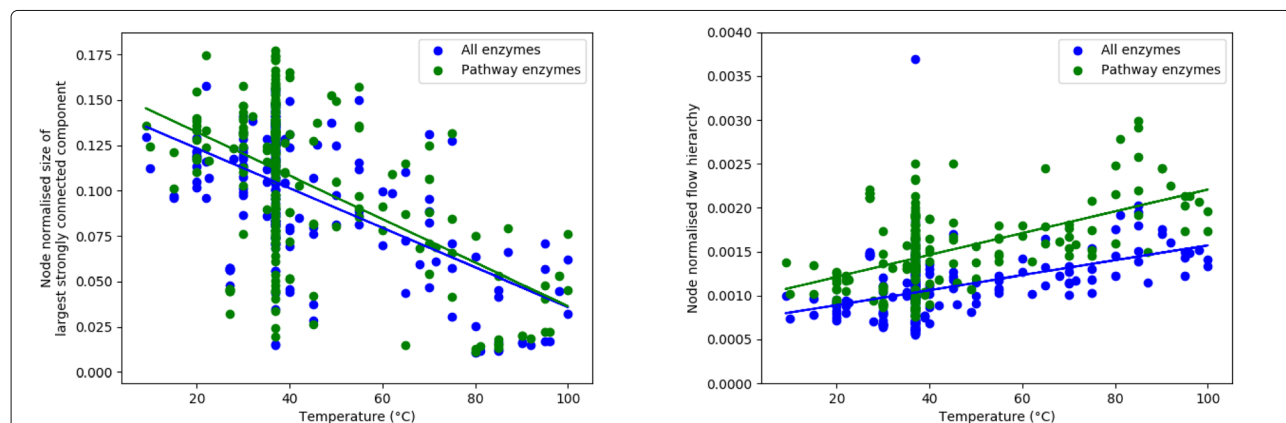
of reactions and can be assessed using the spectrum of the network [18].

More precisely, for each species' graph we computed the adjacency matrix containing all recorded compounds among all species yielding a matrix  $A$  of dimensions  $3194 \times 3194$ . We compute the Laplacian matrix  $\mathcal{L}$  and extract its spectra (the ordered from high to low list of eigenvalues of  $\mathcal{L}$ ). These values must be comprised within the interval  $[0, 2]$ . For example, for a star graph with  $n$  vertices, the eigenvalues are 0, 1 (with multiplicity  $n - 2$ ) and 2, and for the cycle on  $n$  vertices the eigenvalues are  $1 - \cos\left(\frac{2\pi k}{n}\right)$  for  $0 \leq k < n$ .

We computed this spectrum for each species, and estimated the average within the same temperature group (see values in Additional file 2). The results are shown on Fig. 4. The results illustrate clearly that for a temperature class structuration, a higher temperature is associated with lower eigenvalues.

In Fig. 4 we can see steps that can be observed at value 1 and 0. Eigenvalues of 0 have a multiplicity equal to the number of connected components in the graph, but also since we computed the adjacency matrices as the union of all compounds for all species, eigenvalues of 0 also reflect nodes that do not actually exist in a given graph but exist in another (Fig. 2a shows that the largest networks have a maximum of around 1500 nodes whilst Fig. 4 shows 3200 eigenvalues). These non-existent nodes are therefore considered as isolated nodes and counted as a connected component, having an eigenvalue of 0.

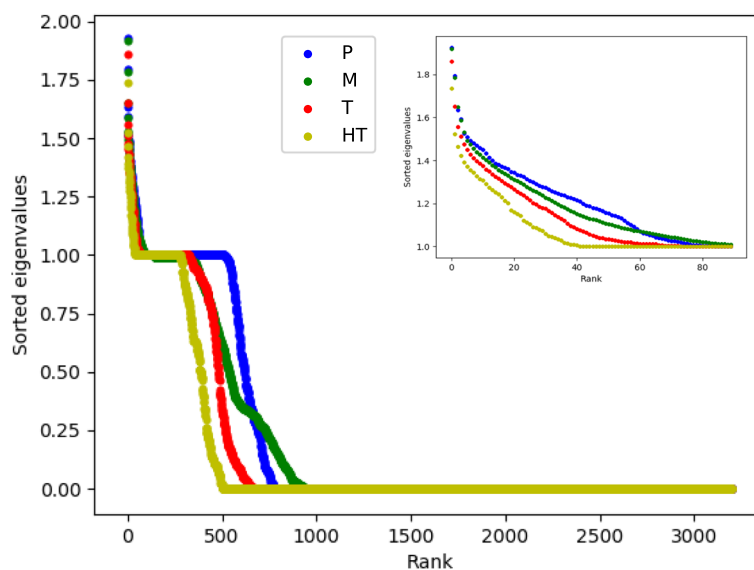
On the other hand, eigenvalues of value 1 with eigenvectors summing to 0 correspond to a particular network pattern :



(a)  $r = -0.56$  with a p-value  $< 10^{-19}$  for the graph with all enzymes,  $r = -0.52$  with a p-value  $< 10^{-16}$  for the graph with only enzymes in known pathways

(b)  $r = 0.44$  with a p-value  $< 10^{-11}$  for all enzymes,  $r = 0.42$ , with a p-value  $< 10^{-10}$  for enzymes in known pathways

**Fig. 3** Robust measures on directed graphs: **a** Pearson's correlation between the node normalised size of the largest strongly connected component and the optimal growth temperature; and **b** Pearson's correlation between the node normalised hierarchy flow and the optimal growth temperature. These measures were applied to directed substrate graphs for all 228 species, without ubiquitous metabolites



**Fig. 4** Averaged sorted eigenvalues of the Laplacian matrix of reaction graph for all groups. In insert a close-up for the first 90 eigenvalues

- Source nodes with at least one target with a single in-degree (whose only predecessor is the source node),
- The said targets.

Source nodes are nodes without predecessors (in-degree of 0).

## Discussion

We have observed that prokaryotic metabolic network properties can correlate with environmental phenotypes, namely with the optimal growth temperature in our study.

First of all, we confirmed the results of [11] that dates more than ten years back: a negative correlation between edge density and optimal growth temperature and between average clustering coefficient and optimal growth temperature, and a positive correlation between maximum likelihood estimate of degree exponent and optimal growth temperature. This clearly shows the robustness of the data from the KEGG database, even though some data has been modified, and some new data has been added. On another hand, the results of our experiments also illustrate the validity of the results of [11]. Although the amount of species in our experiments is doubled compared to the number of prokaryotic species in [11], the trends are still the same.

These results hold for a particular subset of known metabolic data of the species. We consider important to take all available data (enzymes) into account. However, in the case where all enzymes associated to the species in KEGG are taken into account, these results do not hold anymore: they change sign of the correlation. The origins

of these inversions are still unclear, but we believe that it could be due to biases in KEGG for non-mesophilic species (especially for thermophilic species).

We believe a directed network representation is more appropriate to model metabolism, so we looked for directed topological properties that were robust for the different reconstruction protocols. We tested directed graph structural properties related to cycles (largest strongly connected component and flow hierarchy) and to connectivity and flow (Laplacian spectrum).

We found that when there are less nodes involved in the largest strongly connected component (cycle), it is linked to higher optimal growth temperatures, and there are more nodes outside of the cycle that are still part of the weakly connected component, which is the measure of hierarchy flow. When more nodes are involved in the largest strongly connected component, it is linked to lower optimal growth temperatures, and there are less nodes outside the cycle that are still part of the weakly connected component. This is valid no matter if all enzymes are considered or only pathway enzymes.

We explored the most common nodes of the largest strongly connected components, and we found several metabolites such as L-glutamate and L-glutamine (found in 213 out of 228 species), pyruvate (found in 213 from 228 species), phosphoenolpyruvate (209/228), carbamate (208/228) and some others which are molecules involved in the most basic cell metabolism, and that may imply the primordial and basic functions of these metabolic cycles.

Amino acid substitutions are reported to be more deleterious for thermophiles than non-thermophiles [19], implying less variability in enzymes, thus less enzymes,

explaining the negative correlation between the number of nodes in our graphs and optimal growth temperature (Fig. 2a). We can hypothesize that for this very reason, the set of core enzymes and metabolites of the metabolism, which could be represented by the largest strongly connected component, also represents a smaller fraction of nodes because of the greater evolutionary pressure given by temperature. Consequently, as the fraction of nodes in the largest strongly connected component is smaller for thermophiles, the fraction of nodes for the flow hierarchy is larger.

On another hand, we see that Laplacian eigenvalues are higher for prokaryotes that preferentially grow in colder environments, showing more particular patterns of connectivity and flow in their networks.

Other directed graph topological properties were tested, with some having significant correlations with optimal growth temperature for both reconstructions, such as the fraction of nodes with an in-degree of 0 (starting-point nodes, input metabolites) or the fraction with an out-degree of 0 (end-point nodes, output metabolites), having both positive correlations with temperature (data not shown), or also the number of some of the triads among the 16 possible triads in a directed network also show significant correlations, positive and negative (data not shown). All of this shows the clear link between the immediate environment and the metabolism of a given species, and can be looked into in different contexts and environments.

To integrate directed graphs and bring it a step further, an interesting future research avenue would be to study the differences, or complementarities between community graphs and single organism graphs, as well as differences in their largest strongly connected components and other directed structural properties.

An important point has to however be made on the direction of chemical reactions. In this work, we fixed the directions of reactions as found in the KEGG database. However, there might be some chemical reactions happening in the opposite direction than the one fixed in KEGG. We believe that it may be interesting to infer directionality of reactions, since homeostasis is extremely important for organisms, and it is regulated, e.g., by enzymes. There is a need to study this problem, for example through thermodynamics, and it would be promising to study the flux of our graphs, in particular from observational data to investigate the dynamics of the biomass.

## Conclusions

We have reproduced the results of [11], and we state that the results mostly hold even with the evolution of the KEGG database, and even while significantly increasing the number of species in the data set. We have found a positive correlation between the degree exponent

estimate and optimal growth temperature, and a negative correlation between the edge density and the temperature and between average clustering coefficient and the temperature.

We have noticed that when we include all KEGG enzymes we could find for a species into metabolic networks, and not only the enzymes from KEGG pathways, the results do not hold anymore.

We propose three directed graph measures, namely, the size of the largest strongly connected component, the flow hierarchy, and the Laplacian spectrum. We have shown that these measures are robust for all considered graphs, and they correlate respectively negatively, positively and negatively to the optimal growth temperature. In all our experiments, we have observed strong links between environmental phenotypes and graph structure.

We have also developed a pipeline to reconstruct metabolic networks taking into account all enzymes. We compared the results of our pipeline to the state-of-the-art results of [11], and we can state that our pipeline yields very reasonable results.

We are currently investigating how robust the metabolic networks are against structural modifications. Finding causal directions from purely observational data is another open challenge.

## Methods

### Prokaryotic species

Our data set contains 228 prokaryotic species where 100 species are from the database used by [11]. We decided to increase the number of species in our experiments, since the number of bacteria in three growth temperature classes was too small (1 psychrophile, 9 hyperthermophiles, and 9 thermophiles). We added 52 species from the Bacterial Diversity metadatabase (BacDive) [20], chosen according to their growth temperature class. We also added 76 mesophilic species from the Human Pan-Microbe Communities (HPMC) database [21]. We obtained the following distribution over four growth temperature classes (from hot to cold): 19 hyperthermophiles (HT), 35 thermophiles (T), 158 mesophiles (M), and 16 psychrophiles (P). Mesophiles are the most well-studied species, thus biasing databases towards these species and explaining the temperature distribution of our species. Hyperthermophilic species are species whose optimal temperature is above 80°C, thermophilic species are ones with the optimal temperature between 50°C and 70°C, mesophilic species live in the range between 20°C and 45°C, and psychrophilic species prefer an environment between -20°C and 10°C.

### Metabolic network reconstruction

There exist a number of ways to produce metabolic networks from chemical reactions. The nodes of such

a metabolic reconstruction can be metabolites (small molecules, substrates, and products of the enzymes), or enzymes. We have built directed and undirected substrate graphs which are metabolite-centered graphs where each substrate is a node and is linked to each product of a metabolic reaction for a given species. Therefore, edges are enzymatic reactions linking substrates to products. Figure 5 sketches a metabolic reaction, a directed graph, and an undirected metabolic network.

To build the metabolic graphs, we downloaded Ensembl or GenBank cDNA FASTA files for the 228 species. Figure 6 shows the reconstruction procedure. We retrieved gene labels from the FASTA files, see Fig. 6a. We then consulted the Kyoto Encyclopedia of Genes and Genomes (KEGG, [22]) database. With the KEGG code for a species and the gene labels, we found the species gene entries, and we extracted all enzyme commission codes (ECs) if the codes were found in complete form, i.e., no hyphen was present in the code. This step is shown on Fig. 6b. We then extracted all substrates and products from the KEGG enzyme entries (Fig. 6c), and we built directed and undirected substrate graphs, which is illustrated by Fig. 6d and e. We excluded 13 species out of the 113 species from [11] because we could not find gene names in the cDNA FASTA files, or the gene names did not match to the KEGG species code, or the species entry (and code) in KEGG simply did not exist anymore.

A common practice in metabolic network reconstruction is to exclude ubiquitous metabolites to make the network more relevant biologically, and because of the great impact on network structure. There is no strict consensus on ubiquitous metabolites, however, the metabolites used as carriers for transferring electrons and common functional groups are regarded as ubiquitous metabolites [23]. Similarly to [11], we defined 13 ubiquitous metabolites:

H<sub>2</sub>O, ATP, ADP, NAD<sup>+</sup>, NADH, NADPH, CO<sub>2</sub>, ammonia, sulfate, thioredoxin, phosphate, pyrophosphate (PP<sub>i</sub>), and H<sup>+</sup>. We also consider NADP<sup>+</sup> to be a ubiquitous metabolite. All these metabolites do not appear in our graphs.

Note that in [11] they directly downloaded the metabolic pathways of the prokaryotes from the KEGG which are curated networks and are, therefore, different from the networks found with the full enzyme set of the prokaryotes. In order to replicate as accurately as possible their results, we also built graphs without enzymes that do not have an associated KEGG pathway (without PATHWAY field in the KEGG enzyme entry).

Indeed, our main objective is to assess bacterial metabolic systems without any a priori knowledge, and therefore keep as much information as possible, which is why we keep all enzymes that can be deduced from the genome. This means that we have kept most inorganic compounds and generic reactions. We therefore may have less metabolic information regarding some nodes when considering generic reactions, but also more information as more data from the database is considered and as all substrates and products are included (pathways sometimes only show the main reactants and not all of them).

For the directed reconstructed graphs, the default direction of the KEGG reaction was used, which is the direction of the catalytic reaction (substrates and products are specified). It is the direction in which the flow of biomass is expected.

Our networks were reconstructed on April 2019, a description of the species and the networks can be found in Additional file 1.

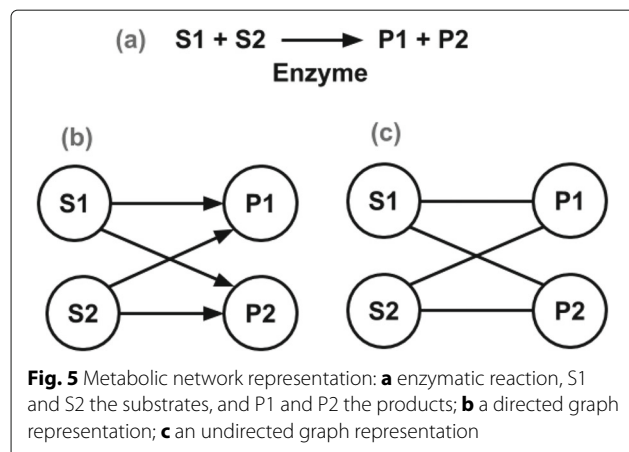
#### Optimal growth temperatures

For the species also considered by [11], we got the optimal growth temperatures from the supplementary material provided with their article. The data originally came from the Prokaryotic Growth Temperature Database (PGTdb) [24]. The access to the PGTdb was not available since we started performing our experiments and later, so, the optimal growth temperature and the growth temperature classes for the rest of the species were taken from the BacDive database [20]. For the species whose optimal growth temperature was given as an interval in the BacDive database, we used the average value of the interval. For the species from the Human Pan-Microbe Communities (HPMC) database, the optimal growth temperature was fixed to 37°C.

#### Measures on directed and undirected graphs

##### Edge density for undirected graphs

Here we use the definition provided in [11] for the edge density for an undirected graph:





```
>AAC73112 cdna chromosome:ASM584v2:Chromosome:190:255:1 gene:b0001 ...
ATGAAACGCATTAGCACCACCATTACCACCACCATTACCATTACCACAGGTAACGGTGCG
GGCTGA

>AAC73113 cdna chromosome:ASM584v2:Chromosome:337:2799:1 gene:b0002 ...
ATGCGAGTGTGGAAAGTTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGGGTTGT
GCGGATATTCGGAAAGCAATGCCAGGCGAGGGGAGGTGGCCACCGTCTCTCTCGCCCCC
GCCAAAATCACCAACCACTGGTGGCGATGATTGAAAAACCAATTAGCGGCCAGGATGCT
...
```

cDNA fasta file

(a) Step 1: From cDNA fasta file, get gene labels.

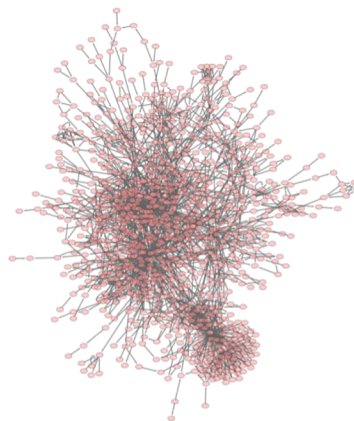
**KEGG gene entries**

**KEGG enzyme entries**

(b) Step 2: From KEGG gene entries, get EC(c) Step 3: From KEGG enzyme entries, get substrates and products.

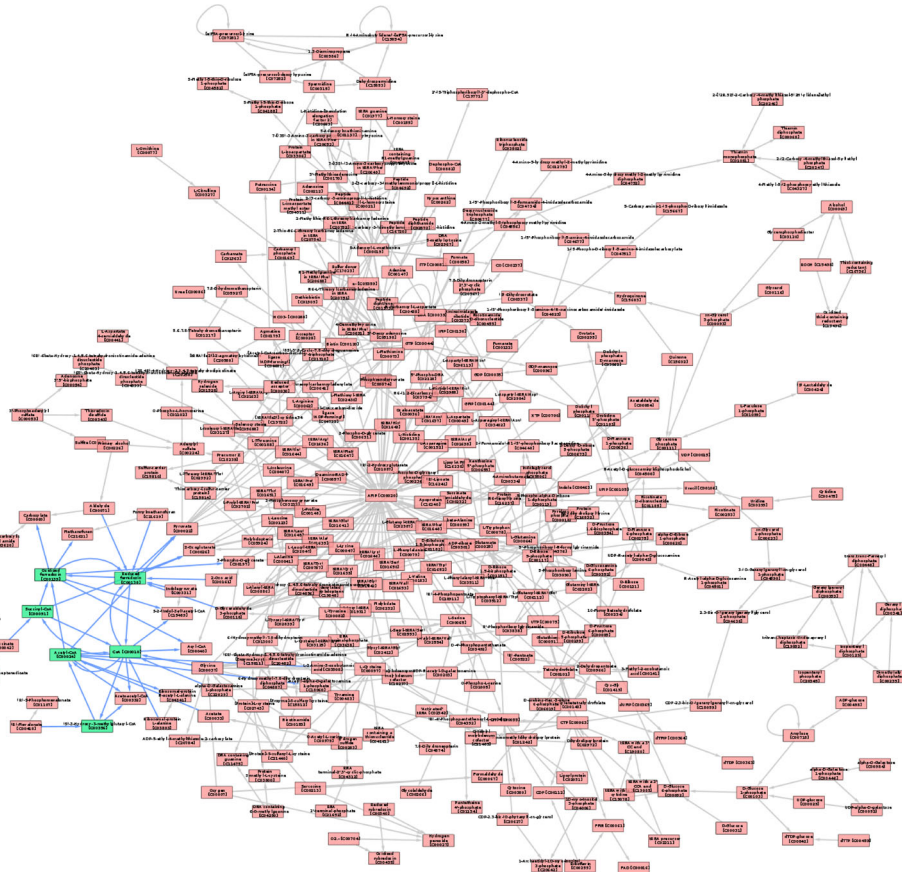


(d) Step 4: Build graph and filter ubiquitous metabolites. Ubiquitous metabolites : H2O, ATP, ADP, NAD+, NADH, NADPH, NADP+, CO2, ammonia, sulfate, thioredoxin, phosphate, PPI, H+.

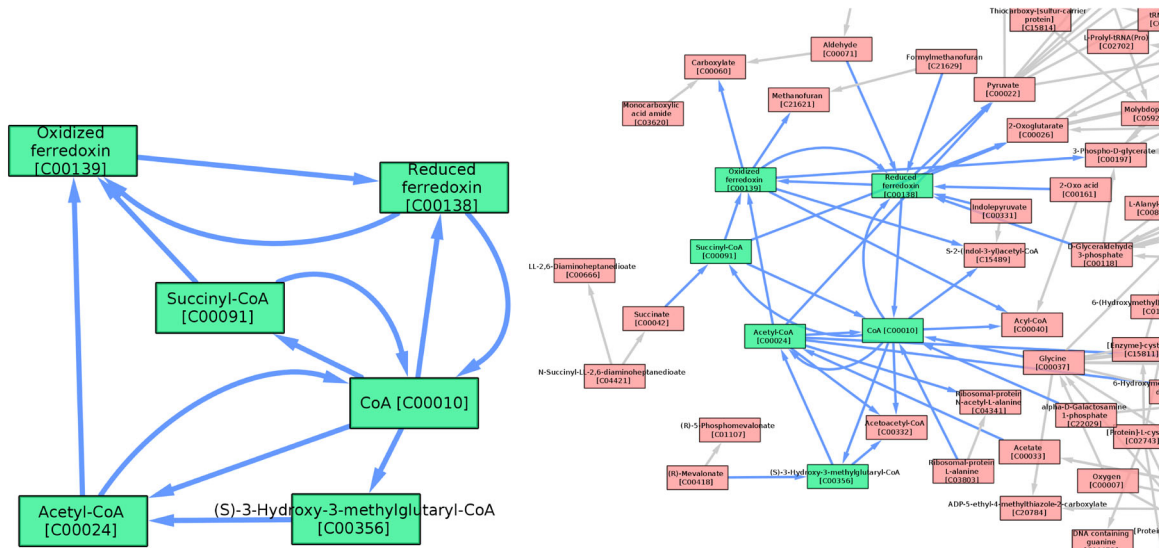


(e) Final graph.

**Fig. 6** Our network reconstruction procedure. **a** Step 1: From cDNA fasta file, get gene labels, **b** Step 2: From KEGG gene entries, get EC codes, **c** Step 3: From KEGG enzyme entries, get substrates and products, **d** Step 4: Build graph and filter ubiquitous metabolites. Ubiquitous metabolites : H2O, ATP, ADP, NAD+, NADH, NADPH, NADP+, CO2, ammonia, sulfate, thioredoxin, phosphate, PPI, H+, **e** Final graph



(a) Weakly connected component containing the largest strongly connected component (nodes in green, edges in blue)



(b) Largest strongly connected component

(c) Zoom of local area containing the largest strongly connected component

**Fig. 7** Visualization of the largest strongly connected component and weakly connected component for *Desulfurococcus amylolyticus* 1221n. **a** Weakly connected component containing the largest strongly connected component (nodes in green, edges in blue), **b** Largest strongly connected component, **c** Zoom of local area containing the largest strongly connected component

$$\text{Edge density} = \frac{E}{N}, \tag{1}$$

where  $E$  is the total number of edges, and  $N$  is the total number of nodes.

**Maximum likelihood estimate of degree exponent**

We follow the definition given by [11]. We assume that the degree distribution  $P(k)$  of our graph follows a power law  $k^{-\gamma}$ . The number of connections  $k$  of a node is called *degree* of a node, and the degree distribution is the degrees of nodes over the whole graph. An estimate via maximum likelihood of the degree exponent  $\gamma$  is as follows:

$$\gamma = 1 + N \times \left[ \sum_{i=1}^N \ln \frac{k_i}{k_{min}} \right]^{-1}, \tag{2}$$

where  $N$  is the number of nodes in the network,  $k_i$  is the degree of node  $i$  and  $k_{min}$  is the smallest degree in the metabolic network. We do not take into account nodes with null degrees for this measure.

**Average clustering coefficient**

Here we have used an approximation of the average clustering coefficient. The local clustering coefficient of a node  $i$  in an undirected graph  $G$  is defined as:

$$C_i = \frac{M_i}{M_{possible}}, \tag{3}$$

where  $M_i$  is the number of triangles formed by a node and two of its neighbours, and  $M_{possible}$  is the number of all possible triangles that could be formed with this node's neighbourhood. The average clustering coefficient corresponds to an average value of local clustering coefficients over all nodes. The approximation we have applied is the one proposed by [25] where the action of choosing a node at random and checking whether its two random neighbours are connected is repeated  $n$  times (we have taken  $n = 1000$ ).

The average clustering coefficient  $\bar{C}$  then becomes:

$$\bar{C} = \frac{M}{n}, \tag{4}$$

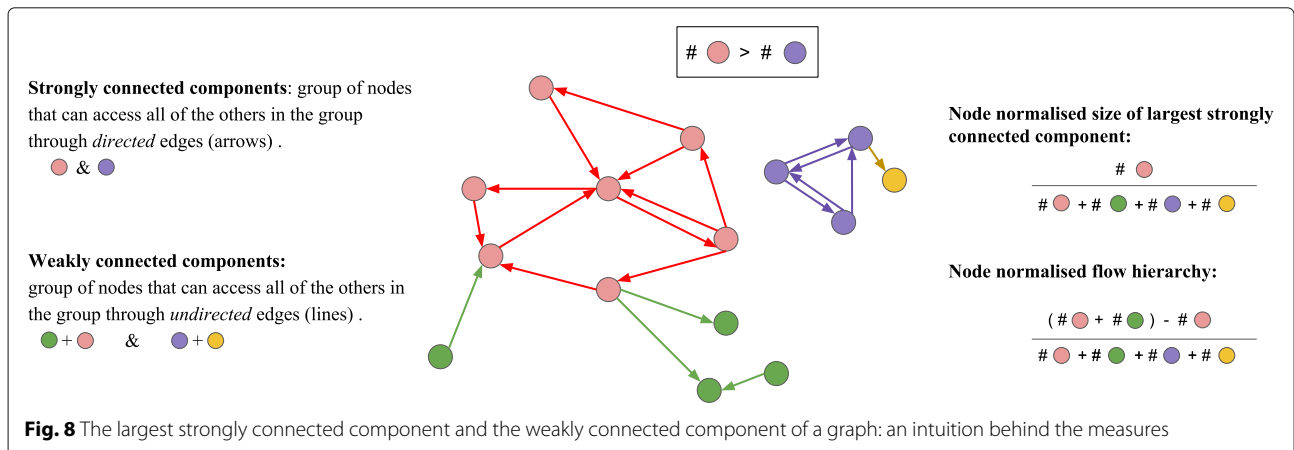
where  $M$  is the number of triangles found, and  $n$  is the number of trials.

**Node-normalised size of the largest strongly connected component**

The largest strongly connected component corresponds to the largest partition of path equivalent nodes in a directed graph. Path equivalence is the property of having a path from node  $v$  to node  $w$ , and a path from  $w$  to  $v$  in a given graph  $G$  [26]. Therefore, the node-normalised size of the largest strongly connected component is the number of nodes of the largest strongly connected component divided by the number of nodes. Note that applying this definition, the strongly connected components are cycles. An example of the largest strongly connected component for *Desulfurococcus amylolyticus 1221n* is shown on Fig. 7. The number of nodes in the largest strongly connected component might be small compared to the number of all nodes in a graph (in the example it is 6 out of 340 nodes).

**Node-normalised hierarchy flow**

A weakly connected component is also a property of directed graphs. It is defined as a group of nodes where each node  $v$  and  $w$  are connected via an undirected path. We have defined the concept of flow hierarchy as the number of nodes that do not participate in the largest strongly connected component. Hierarchy flow can therefore be deduced from the subtraction of the strongly connected component from the weakly connected component. To be precise, we take the largest strongly connected component and the weakly connected component containing it, and we then divide the remainder nodes by the number of nodes in the graph (normalization). This procedure is



drafted on Fig. 8, and can be observed in the example of *Desulfurococcus amylolyticus* 1221n in Fig. 7.

### Laplacian matrix and spectrum

We collected the set of all compounds described in the different graphs to create a standardized adjacency matrix  $A$  for each species indexed by these vertices  $A_{xy} = 1$  if a *directed* edge exists from  $x$  to  $y$ . By construction,  $A$  is not usually symmetric. We compute the Laplacian matrix  $\mathcal{L}$  also indexed by the vertices whose sum over the columns are equal to zero and  $L_{xy} = -A_{xy}$  if  $x \neq y$ . We computed the spectrum—the list of eigenvalues—of  $\mathcal{L}$  and ordered it by highest to lowest. We computed the average of this sorted vector for all species within a temperature class.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12859-019-3112-y>.

Code available at <https://github.com/AWebZen/micrometabolicnetworks>.

**Additional file 1:** This file contains our prokaryotes and structural properties information in an XLSX format. It contains the species name, KEGG species code, optimal growth temperature, temperature class, number of nodes for all enzyme graphs, number of nodes for pathway graphs, the number of edges for all enzyme graph, number of edges for the pathway graph, the average clustering coefficient for the all enzyme graphs, the average clustering coefficient for the pathway graphs, the degree exponent for the all enzyme graphs, the degree exponent for the pathway graphs, the edge density for the all enzyme graphs, the edge density for the pathway graphs, the node normalised flow hierarchy for the all enzyme graphs, the node normalised flow hierarchy for the pathway graphs, the node normalised size of the largest strongly connected component for the all enzyme graphs, the node normalised size of the largest strongly connected component for the pathway graphs.

**Additional file 2:** This file contains the average Laplacian spectrum per temperature class, sorted in decreasing order. The Laplacian matrix size is the number of nodes in the union of nodes for all our networks.

### Abbreviations

ADP: Adenosine Diphosphate; ATP: Adenosine Triphosphate; BacDive: Bacterial Diversity metadatabase; cDNA: Complementary Deoxyribonucleic Acid; CO<sub>2</sub>: Carbon dioxide; CoA: Coenzyme A; EC: Enzyme Commission number; FBA: Flux Balance Analysis; H<sup>+</sup>: Proton; H<sub>2</sub>O: Water; HPMC: Human Pan-Microbe Communities database; HT: Hyperthermophiles; KEGG: Kyoto Encyclopedia of Genes and Genomes database; M: Mesophiles; NAD<sup>+</sup>: Oxidized Nicotinamide Adenine Dinucleotide; NADH: Reduced Nicotinamide Adenine Dinucleotide; NADP<sup>+</sup>: Oxidized Nicotinamide Adenine Dinucleotide Phosphate; NADPH: Reduced Nicotinamide Adenine Dinucleotide Phosphate; P: Psychrophiles; PGTdb: Prokaryotic Growth Temperature database; PP: Pyrophosphate; T: Thermophiles

### Acknowledgments

The authors acknowledge and thank the anonymous reviewers for their suggestions that allowed the improvement of our manuscript.

### Authors' contributions

AWZ, NS, and HAS conceived and performed the research. AWZ, NS, and HAS wrote the manuscript. All authors read and approved its final version.

### Funding

This work was supported by AAP Inserm/CNRS Défi Santé numérique *Modelling metabolism of intestinal microbiome by multi-omics statistical data integration*.

### Availability of data and material

The data used in the numerical experiments are publicly available. KEGG database is available at <https://www.kegg.jp/>, BacDive database can be accessed at <https://bacdive.dsmz.de/>, and HPMC database at <http://www.hpmcd.org/>. PGT database is no longer available, growth temperatures were taken from [11] additional file.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 26 October 2018 Accepted: 20 September 2019

Published online: 15 October 2019

### References

- Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* 2002;30(1):42–6.
- Ebenhöh O, Handorf T, Heinrich R. A cross species comparison of metabolic network functions. *Genome Inform.* 2005;16(1):203–5.
- Janga SC, Babu MM. Network-based approaches for linking metabolism with environment. *Genome Biol.* 2008. <https://doi.org/10.1186/gb-2008-9-11-239>.
- Pah AR, Guimerà R, Mustoe AM, Amaral LAN. Use of a global metabolic network to curate organismal metabolic networks. *Sci Rep.* 2013. <https://doi.org/10.1038/srep01695>.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L. The large-scale organization of metabolic networks. *Nature.* 2000;407(6804):651.
- Ma H, Zeng A-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics.* 2003;19(2):270–7.
- Zhu D, Qin ZS. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics.* 2005;6(1):8.
- Oberhardt MA, Chavali AK, Papin JA. Flux balance analysis: interrogating genome-scale metabolic networks. *Methods Mol Biol.* 2009;500:61–80.
- Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet.* 2005;37(12):1372–5.
- Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc Natl Acad Sci USA.* 2008;105(38):14482–7.
- Takemoto K, Nacher JC, Akutsu T. Correlation between structure and temperature in prokaryotic metabolic networks. *BMC Bioinformatics.* 2007;8(1):303. <https://doi.org/10.1186/1471-2105-8-303>.
- Pearcy N, Crofts JJ, Chuzhanova N. Network motif frequency vectors reveal evolving metabolic network organisation. *Molecular Biosyst.* 2015;11(1):77–85. <https://doi.org/10.1039/c4mb00430b>.
- Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief Bioinform.* 2006;8(3):243–55.
- Zhou W, Nakhleh L. Properties of metabolic graphs: biological organization or representation artifacts?. *BMC Bioinformatics.* 2011;12(1):132.
- Wunderlich Z, Mirny LA. Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J.* 2006;91:2304–11.
- Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG. Using graph theory to analyze biological networks. *BioData Min.* 2011;4(1):10.
- Jonnalagadda S, Srinivasan R. An efficient graph theory based method to identify every minimal reaction set in a metabolic network. *BMC Syst Biol.* 2014;8(1):28.
- Doob M. Eigenvalues of graphs. *Top Algebraic Graph Theory.* 2004;102:30.
- Friedman R, Drake JW, Hughes AL. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics.* 2004;167(3):1507–12. <https://doi.org/10.1534/genetics.104.026344>.
- Söhngen C, Podstawka A, Bunk B, Gleim D, Vetcinova A, Reimer LC, Ebeling C, Pendarovski C, Overmann J. Bacdiv— the bacterial diversity

metadatabase in 2016. *Nucleic Acids Res.* 2016;44(D1):581–5. <https://doi.org/10.1093/nar/gkv983>.

21. Forster SC, Browne HP, Kumar N, Hunt M, Denise H, Mitchell A, Finn RD, Lawley TD. Hpmcd: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res.* 2016;44(D1):604–9. <https://doi.org/10.1093/nar/gkv1216>.
22. Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
23. Altaf-Ul-Amin M, Katsuragi T, Sato T, Kanaya S. A Glimpse to Background and Characteristics of Major Molecular Biological Networks: BioMed Research International; 2015.
24. Huang SL, Wu LC, Laing HK, Pan KT, Horng JT. Pgtdb: a database providing growth temperatures of prokaryotes. *Bioinformatics.* 2004;20: <https://doi.org/10.1093/bioinformatics/btg403>.
25. Schank T, Wagner D. Approximating clustering-coefficient and transitivity. Karlsruhe: Universität Karlsruhe; 2004.
26. Nuutila E, Soisalon-Soininen E. On finding the strongly connected components in a directed graph. *Inf Process Lett.* 1994;49(1):9–14. [https://doi.org/10.1016/0020-0190\(94\)90047-7](https://doi.org/10.1016/0020-0190(94)90047-7).

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

