Supplementary Material for:
Network motif frequency vectors reveal evolving metabolic network organisation

Contents

1 Relation between environmental conditions and basic structural properties of the metabolic networks
   1.1 Average total degree .............................................. 1
   1.2 Average path length ............................................ 3
   1.3 Clustering coefficient ........................................... 4
   1.4 Spearman’s partial correlation between $P$ and environment conditioned on basic network measures ................. 4

2 Motif dictionary
   2.1 3-node subgraphs ................................................. 6
   2.2 4-node subgraphs ................................................. 7

3 Significant metabolites: Motif 9
   3.1 Habitat variability .............................................. 15
   3.2 Oxygen requirement ............................................ 15

1 Relation between environmental conditions and basic structural properties of the metabolic networks

1.1 Average total degree

The total degree of a node is the total number of arcs (directed edges) connected to that node, where the arc can be pointing in either direction (i.e. we sum the in-coming and out-going arcs to and from that node). The average total degree $d_{tot}^j$ for the $j$th network is then calculated by taking the mean over all nodes in the network.

Habitat variability

Figure S1 (a) shows that the trend for the average total degree is to increase with variability in the environment, although not monotonically. In Figure S1 (b) we plot the average total degree against our global significance score $P_{\text{global}}$ and find a significant correlation ($r = 0.8983$, $p < 10^{-5}$).
Figure S1: (a) Relationship between environmental variability and the average total degree for the 6 environmental classes. Note that here we plot the mean value over each environmental class: Obligate, Specialised, Aquatic, Facultative, Multiple and Terrestrial. (b) The average total degree plotted against the global significance score $P_{global}$ for the 115 bacterial networks.

Figure S2: (a) Relationship between growth conditions, specifically oxygen requirements, and the average total degree. Note that here we plot the mean value over each class and error bars represent standard errors. (b) The average total degree plotted against the global significance score $P_{global}$ for the 383 bacterial networks.
**Oxygen requirements**

Figure S2 (a) shows that the average total degree does not appear to follow any particular trend regarding oxygen requirements, despite being correlated to the global significance score $P_{\text{global}}$ ($r = 0.7932$, $p < 10^{-5}$) (Figure S2 (b)).

### 1.2 Average path length

The average path length $L$ is defined as the average shortest path length between all pairs of nodes in the network. $L$ is measure of how efficient information can be transported across the network.

**Habitat variability**

Figure S3 (a) shows that the average path length is the smallest for the networks within the obligate class (i.e. the most specialised) and the largest for the terrestrial class (i.e. the most varied). The specialised, aquatic, facultative and multiple class however, all have very similar values. In Figure S3 (b) we have plotted the average path length against the global significance score $P_{\text{global}}$ and find a significant correlation ($r = 0.8723$, $p < 10^{-5}$).

![Figure S3: (a) Relationship between environmental variability and the average path length for the 6 environmental classes. Note that here we plot the mean value over each environmental class: Obligate, Specialised, Aquatic, Facultative, Multiple and Terrestrial. (b) The average path length plotted against the global significance score $P_{\text{global}}$ for the 115 bacterial networks.](image)

**Oxygen requirements**

Figure S4 (a) shows that the mean average path length is larger for the metabolic networks that evolved in the presence of oxygen, that is, the aerobic and facultative class. In Figure S4 (b) we have plotted the average path length against the global significance score $P_{\text{global}}$ for the 383 metabolic networks and find a significant correlation ($r = 0.8617$, $p < 10^{-5}$).
Oxygen Requirements

Figure S4: (a) Relationship between growth conditions, specifically oxygen requirements, and the average path length. Note that here we plot the mean value over each class and error bars represent standard errors. (b) The average path length plotted against the global significance score $P_{\text{global}}$ for the 383 bacterial networks.

1.3 Clustering coefficient

The clustering coefficient measures the extent to which the nodes in the network tend to cluster. More formally, the clustering coefficient is defined as the fraction of a node’s neighbours that are also neighbours of each other. We compute the average clustering coefficient by taking the mean value over all nodes within the network.

Habitat variability

Figure S5 (a) shows that the clustering coefficient does not follow any particular trend as regards environmental habitat. Figure S5 (b) shows that the average clustering coefficient is only weakly correlated with the global significance score $P_{\text{global}}$ ($r = 0.2531$, $p < 0.01$).

Oxygen requirements

Figure S6 (a) shows a relationship similar to the average total degree (Figure S2), that is, the facultative class has a significantly larger amount of clustering present than the aerobic and anaerobic classes. Figure S6 (b) shows that the average clustering coefficient and the global significance score $P_{\text{global}}$ are not correlated ($r = 0.0737$, $p = 0.1489$).

See [2] for more details concerning the network measures above.

1.4 Spearman’s partial correlation between $P$ and environment conditioned on basic network measures

Spearman’s partial correlation between $X$ and $Y$ conditioned on $Z$ allows one to compute the correlation between $X$ and $Y$, discounting the correlations between $X$ and $Z$ and between $Y$
Figure S5: (a) Relationship between environmental variability and the average clustering coefficient for the 6 environmental classes. Note that here we plot the mean value over each environmental class: Obligate, Specialised, AQuatic, Facultative, Multiple and Terrestrial. (b) The average clustering coefficient plotted against the global significance score $P_{\text{global}}$ for the 115 bacterial networks.

Figure S6: (a) Relationship between growth conditions, specifically oxygen requirements, and the average clustering coefficient. Note that here we plot the mean value over each class and error bars represent standard errors. (b) The average clustering coefficient plotted against the global significance score $P_{\text{global}}$ for the 383 bacterial networks.
and $Z$. We computed the correlation between the global motif significance score and both variability and oxygen requirements conditioned on the simpler network metrics considered in the previous section (degree, path-length and clustering) and found that our results remained significant ($c = 1$, $p < 10^{-3}$ in all instances). Note that we use Spearman's correlation since the data consists of a mixture of both ordinal and continuous values; correlations were computed using the `partialcorr` function which is available in the MATLAB Statistics Toolbox.

## 2 Motif dictionary

The motif dictionary provides a graphical description of the 13 3-node subgraphs and the 199 4-node subgraphs used in this study. Here, the top label corresponds to a motif's identification number when using the `mfinder` software (http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html), whereas the bottom labels correspond to the motif numbering used in this work.

### 2.1 3-node subgraphs
2.2 4-node subgraphs

id14 motif 14

id28 motif 15

id30 motif 16

id74 motif 17

id76 motif 18

id78 motif 19

id90 motif 20

id92 motif 21

id94 motif 22

id204 motif 23

id206 motif 24

id222 motif 25

id280 motif 26

id282 motif 27

id286 motif 28

id328 motif 29

id330 motif 30

id332 motif 31

id334 motif 32

id344 motif 33

id346 motif 34

id348 motif 35

id350 motif 36

id390 motif 37
3 Significant metabolites: Motif 9

3.1 Habitat variability

Figure S7 shows the mean frequency for metabolites occurring within motif 9 for the 115 metabolic networks, grouped into the specialised (blue bars) and varied (red bars) classes. Here metabolites are displayed in decreasing order according to the varied class. Figure S7 shows the 54 metabolites that were found at least once across the 115 metabolic networks. We find that the distribution of metabolites is slightly broader for the varied class, similar, but less prominent, to the results obtained for motif 5. Using Chi-square tests (Fisher’s Exact test) we explored group differences for the individual metabolites. Figure S8 identifies only one metabolite, RNA, for which significant differences were found (Fisher’s Exact test, $p < 0.001$).

3.2 Oxygen requirement

Figure S9 shows the mean frequency for metabolites occurring within motif 9 for the 383 metabolic networks that evolved in either the presence or absence of oxygen. Here metabolites are displayed in decreasing order according to the aerobic class (blue bars). Figure S9 shows the 65 metabolites that were found at least once across the two classes. Note that the distribution for the aerobic class and anaerobic class for motif 9 are a lot closer than was obtained for motif 5. Figure S10 shows that the metabolites with the most significant differences (Fisher’s Exact test, $p < 0.001$) included Glutathione, L-Arginine, L-Citrulline, N-(L-Arginino)succinate, Succinate, Succinyl-CoA and O-Succinyl-L-homoserine.
Figure S7: Mean normalised frequency for the 54 metabolites obtained for the 115 metabolic networks. Blue bars represent the specialised class and the red bars represent the varied class. Here, the metabolites are in descending order of the metabolite frequencies for the varied class.

Figure S8: Mean normalised frequency for the 54 metabolites obtained for the 115 metabolic networks. Error bars are standard errors. Asterisks indicate levels of significance, with *, **, and *** corresponding to $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively. Metabolite names are provided for the most significant metabolites.
Figure S9: Mean normalised frequency for the 65 metabolites obtained for the 383 metabolic networks. Blue bars represent the aerobic-facultative class and the red bars represent the anaerobic class. Here, the metabolites are in descending order of the metabolite frequencies for the aerobic and facultative class.

Figure S10: Mean normalised frequency for the 65 metabolites obtained for the 383 metabolic networks. Error bars are standard errors. Asterisks indicate levels of significance, with *, **, and *** corresponding to $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively. Metabolite names are provided for the most significant metabolites.
References
